Mariana Cristina Gonçalves Mourão

Licenciada em Engenharia Biomédica

Modelação e Predição do Valor do tempo de vida do cliente (*Customer Lifetime Value*)

Dissertação para obtenção do Grau de Mestre em

Análise e Engenharia de Big Data

Orientadora: Regina Bispo, Professora Auxiliar,
Universidade Nova de Lisboa,
Faculdade de Ciências e Tecnologia

Júri

Presidente: Doutor Pedro Barahona
Arguentes: Doutor José Costa da Cruz
Doutora Regina Bispo

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Fevereiro, 2021

Modelação e Predição do Valor do tempo de vida do cliente (*Customer Lifetime Value*)

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Aos meus pais.

Agradecimentos

Obrigada Faculdade de Ciências e Tecnologias da Universidade Nova de Lisboa. Obrigado pelos dois anos intensos e desafiadores que aqui passei, pelos conhecimentos transmitidos e, acima de tudo pelo carinho com que me acolheste fazendo-me sentir em casa mesmo estando tão longe.

Começo por agradecer à minha orientadora, Professora Regina Bispo pelo apoio, conselhos e ensinamentos transmitidos ao longo do desenvolvimento deste trabalho. Sem si, não teria sido possível crescer tanto quanto cresci durante este ano. Levo na bagagem, graças a si, ferramentas que me permitiram desenvolver-me profissionalmente e pessoalmente.

Aos meus pais, Maria e António, tenho a agradecer tudo aquilo que eu sou hoje, foram vocês que me ensinaram a lutar por aquilo que quero e que é os desejos servem para ser concretizados. Foram vocês que sempre me deram conselhos importantes e tiveram paciência para me acalmar nos momentos mais difíceis. Este projeto só foi possível de concretizar porque tenho na minha vida pessoas incríveis, como vocês, que inspiram todos os dias.

Ao meu irmão, Fábio, tenho que agradecer o otimismo com que me ensina a lidar com a vida. Apesar de seres o meu irmão mais novo tenho muito a aprender contigo.

Obrigada às minhas amigas, que apesar da distância física, estão lá sempre para me dar os melhores conselhos e/ou abraços.

Agradeço à equipa de Análise e Modelos do Millennium BCP, em primeiro lugar pela disponibilização de dados, imprescindíveis à realização desta dissertação e, em segundo lugar pela forma como me acolheram na vossa equipa, fazendo com que me sentisse em casa, pelos conhecimentos transmitidos diariamente e pelo espírito de desafio e aperfeiçoamento constante que me transmitem. Um agradecimento especial à Amélia Goulão pela compreensão e conselhos dados durante a realização deste projeto; e, ao Miguel Aguiar e ao Pedro Bandeira Marques pela paciência que tiveram em ensinar-me SAS. Sem vocês tudo teria sido muito mais difícil.

Por último, um agradecimento especial ao João pelo apoio constante, por me tornares todos os dias uma pessoa melhor e por seres o meu porto de abrigo.

Que as próximas aventuras e desafios continuem a ser partilhados com todos vocês.

Mariana

Resumo

O conceito "valor de tempo de vida do cliente", *Customer Lifetime Value* (CLV) da literatura anglo-saxónica, surgiu devido à necessidade de ao longo dos anos, as empresas reterem os clientes mais lucrativos. Conhecendo o valor de cada um dos seus clientes, as empresas conseguem alocar os seus recursos de forma mais consciente, bem como determinar o máximo de investimento que é viável fazer em cada cliente para o conseguir reter.
Ao longo dos anos têm sido desenvolvidos vários métodos para a medição do CLV como, por exemplo, as abordagens RFM (*Recency, Frequency and Monetetary Value*), em que o valor do cliente é calculado tendo por base as variáveis recência, frequência e valor monetário da compra, ou SOW (*Share-of-Wallet*) que consiste na simples segmentação dos clientes como "bons" ou "maus". Este tipo de soluções, mais antigas, tem como principal desvantagem o facto de segmentarem os clientes, tendo apenas em conta a contribuição passada do cliente. Para ultrapassar este problema alguns autores propuseram, mais recentemente, a utilização de modelos de *Machine Learning* e *frameworks* de *big data* conseguindo obter uma maior precisão na previsão do CLV.
Os objetivos deste estudo são a construção de modelos preditivos do CLV.
Neste estudo, numa primeira fase, foi feita a revisão de literatura sobre o conceito de CLV, bem como sobre as técnicas usadas para o calcular e modelar. Seguiu-se a implementação de modelos de *Machine Learning* nomeadamente dos modelos *Classification and Regression Trees* (CART), *Random Forest* (RF), Cadeias de Markov, *Multi Layer Perceptron* (MLP) e algoritmos de *clustering*. No final fez-se uma comparação entre esses métodos e caracterização de cada um dos grupos resultante da análise de *clusters*.
O *dataset* usado provêm de uma instituição bancária portuguesa, incluindo variáveis demográficas e sócio-económicas. A partir do conjunto inicial de variáveis foi feita a construção de novas variáveis, que foi sempre apoiada na revisão da literatura e simultaneamente ajustada ao negócio bancário.
Conclui-se que, as RF são o modelo que apresenta maior eficácia na previsão do CLV, registando um erro percentual de 5.98%.
Neste estudo, foi também realizada uma análise de *clusters* com o propósito de obter um melhor conhecimento dos produtos que suscitam interesse a cada grupo de clientes e consequentemente servir de base para desenhar campanhas de *marketing* personalizadas. Em particular, foram encontrados 5 *clusters* de clientes interesses e valores de CLV distintos.

Palavras-chave: Customer Lifetime Value (CLV), Banca, Rentabilidade, Machine Learning, Clustering, Árvores de Decisão, Cadeias de Markov, Multi layer Perceptron (MLP)

Abstract

The concept of Customer Lifetime Value (CLV) comes from Anglo-Saxon literature and started emerging when companies felt the need to retain the most profitable customers. Therefore when a company knows the value of each of their customers, they are able to allocate their resources more effectively and determine the maximum investment that is feasible to make in each customer to retain it.

Over the years various methods have been developed to measure CLV, such as the RFM (Recency, Frequency and Monetary Value) approach, where the customer value is calculated based on the variables recency, frequency, and monetary value, or SOW (Share-of-Wallet) which consists of simply segmenting customers as "good"or "bad". Comparatively, in this older type of solution, the main disadvantage is the fact that they divide customers, taking into account only the past contribution of the customer. In order to overcome this problem some authors have proposed, more recently to use Machine learning models and big data frameworks achieving greater accuracy in CLV forecast.

The objectives of this study is the construction of predictive models of CLV.

In this study, we start with a literature review on the concept of CLV as well as the techniques used to calculate and model it, followed by the implementation of the Machine Learning models, in particular Classification and Regression Trees (CART), Random Forests (RF), Markov Chains, Multi-Layer Perceptron (MLP), and clustering algorithms. Finally, we make a comparison between these methods and characterize each of them into groups resulting from cluster analysis.

The dataset used comes from a Portuguese banking institution and includes demographic and sociology-economic variables, from the initial set of variables we constructed new variables, which has always been supported in the literature review and simultaneously adjusted to the banking business.

We concluded the Random Forest (RF) model is the most effective in forecasting the CLV, recording a percentage error of 5.98%. Finally, we analyzed the cluster with the intent of better understanding the product interest of each group, thus building more personalized marketing campaigns bearing in mind the CLV value for each group. We found 5 clusters of clients with different CLV values and interests.

Keywords: Customer Lifetime Value(CLV), Bank, Profitable, Machine Learning, Clustering, Decision Tree, Markov Chain, Multilayer Perceptron (MLP)

Índice

Lista de Figuras **xv**

Lista de Tabelas **xvii**

Glossário **xix**

Siglas **xxiii**

1 Introdução **1**
1.1 Definição . . . 1
1.2 Objetivos . . . 2
1.3 Motivação . . . 4
1.4 Contextualização . . . 5

2 Estado da arte **7**
2.1 Cálculo do *CLV* . . . 7
2.1.1 *Twelve Month-Prune* . . . 7
2.1.2 RFM . . . 8
2.1.3 Métodos determinísticos . . . 9
2.2 Modelação do *CLV* . . . 13
2.2.1 Métodos Estatísticos . . . 14
2.2.2 *Machine Learning* . . . 15
2.2.3 *Big Data* . . . 18
2.2.4 Outras Abordagens . . . 19

3 *Dataset* **21**
3.1 Descrição e caracterização do *dataset* . . . 21
3.2 Pré-processamento dos dados . . . 22

4 Métodos **31**
4.1 Análise exploratória e descritiva do *dataset* . . . 31
4.2 CART . . . 32
4.3 Random Forest . . . 36
4.4 Cadeias de Markov . . . 37

4.5 *Multilayer Perceptrons* . . . 38
4.6 Avaliação dos modelos supervisionados . . . 41
4.7 *K-Means* . . . 41

5 Resultados **43**
5.1 Análise exploratória e descritiva do *dataset* . . . 43
5.1.1 Variáveis demográficas . . . 51
5.1.2 Variáveis transaccionais . . . 55
5.1.3 Variáveis de fidelização . . . 65
5.1.4 Variáveis de posse . . . 68
5.1.5 Variáveis de reclamação . . . 90
5.1.6 CLV . . . 92
5.2 Modelos . . . 94
5.2.1 CART . . . 94
5.2.2 Random Forest . . . 100
5.2.3 Cadeias de Markov . . . 104
5.2.4 *Multilayer Perceptron* . . . 112
5.2.5 *K-means* . . . 115

6 Conclusão **121**

Referências **123**

Apêndices **129**

A Cadeias de Markov: Definição dos estados utilizando a rentabilidade do ano anterior **129**

Lista de Figuras

1.1 Distribuição de padrões de compra por cliente [57]. 3
1.2 *Timeline* da dissertação 3

5.1 Variação do CLV (em euros) em função do grau de instrução (ver legenda na tabela 5.5) 52
5.2 Distribuição do CLV(euros/mês) em função do sexo. (F - Feminino; M-Masculino). 53
5.3 Variação do CLV (em euros/mês) em função do estado civil (ver legenda na tabela 5.8) 54
5.4 Distribuição do valor total de depósitos (euros/mês) 57
5.5 Distribuição do valor de depósitos em ATM (euros/mês) 58
5.6 Distribuição do valor do depósitos no balcão depósitos (euros/mês). 59
5.7 Distribuição do número de compras a crédito. 60
5.8 Distribuição dos valores de compras a crédito (euros/mês) 61
5.9 Distribuição do número de transferências a crédito (euro/mês). 62
5.10 Distribuição do valor das transferências a crédito (euros/mês) 63
5.11 Distribuição do número de transações por iniciativa própria (por mês). . . . 64
5.12 Antiguidade dos clientes (anos). 66
5.13 Distribuição do valor das simulações (euros/mês). 67
5.14 Distribuição do montante em produtos (euros/mês). 69
5.15 Distribuição do valor dos recursos a título (euros/mês). 70
5.16 Distribuição do valor dos recursos a prazo (euros/mês). 71
5.17 Distribuição do valor de recursos à ordem (euros/mês) 72
5.18 Distribuição do valor de crédito vivo (euros/mês). 73
5.19 Distribuição do valor de crédito à habitação (euros/mês). 74
5.20 Distribuição da densidade do valor de crédito pessoal (euros/mês). 75
5.21 Distribuição do valor do seguro multirriscos (euros/mês). 76
5.22 Distribuição do valor do seguro de acidentes pessoais (euros/mês). 77
5.23 Distribuição do valor do seguro automóvel (euros/mês). 78
5.24 Distribuição valor do seguro de saúde (euros/mês). 79
5.25 Distribuição do valor de seguro de vida (euros/mês). 80
5.26 Distribuição do valor seguro de risco não vida (euros/mês). 81
5.27 Distribuição do valor da poupança reforma (euros/mês) 83

5.28 Distribuição do valor do património financeiro (euros). 84
5.29 Distribuição do valor da rentabilidade líquida dos últimos 12 meses (anos): valores positivos. 85
5.30 Distribuição do valor da rentabilidade líquida dos últimos 12 meses (anos): valores negativos. 86
5.31 Distribuição do valor da rentabilidade líquida há 2 anos antes (euros): valores positivos. 87
5.32 Distribuição do valor da rentabilidade líquida há 2 anos antes (euros): valores negativos. 88
5.33 Distribuição do valor do saldo da conta corrente (euros/mês) 89
5.34 Distribuição do valor estornado (euros/mês). 91
5.35 Distribuição do CLV (euros): valores positivos. 92
5.36 Distribuição do CLV (euros): valores negativos. 93
5.37 Representação do modelo CART gerado utilizando todas as variáveis como *input* (n - número de clientes contidos na folha). 94
5.38 Variação do erro absoluto médio (*Mean Absolute Error* (MAE)) em função do CLV. 95
5.39 Variação do erro percentual em função do CLV. 96
5.40 Ganho de informação associado a cada variável. 97
5.41 Modelo CART gerado utilizando como *input* as variáveis que apresentam um ganho de informação superior a 3% (n - número de clientes contidos na folha). 98
5.42 MAE por decil 99
5.43 Variação do erro absoluto médio (MAE) em função do CLV. 99
5.44 Variação do erro quadrático médio em função do número de árvores consideradas 100
5.45 Variação do erro absoluto médio(MAE) em função do CLV. 101
5.46 Erro percentual por decil. 101
5.47 %IncMSE de cada variável. 102
5.48 MAE por decil. 103
5.49 Erro percentual por decil. 103
5.50 Modelo CART 1 (n - Número de clientes alocados à folha). 106
5.51 Modelo CART 5 (n - Número de clientes alocados à folha). 107
5.52 Modelo CART 6 (n - Número de clientes alocados à folha). 108
5.53 Representação gráfica da matriz de transição 110
5.54 Relação entre o número de *clusters* e a variabilidade intra-*clusters* (a linha azul assinala a solução para $k = 5$) 115

A.1 Modelo CART 12 (n - Número de clientes alocados à folha). 130
A.2 Modelo CART 13 (n - Número de clientes alocados à folha). 131
A.3 Modelo CART 16 (n - Número de clientes alocados à folha). 132

Lista de Tabelas

3.1 Descrição das variáveis 23
3.2 Variáveis criadas 28

5.1 Análise descritiva das variáveis discretas 44
5.2 Análise descritiva das variáveis contínuas (em euros/mês) 46
5.3 Análise descritiva das variáveis construídas nesta dissertação 48
5.4 Análise descritiva das variáveis de rentabilidade 50
5.5 Distribuição dos clientes por grau de instrução 51
5.6 Distribuição dos clientes por profissão 52
5.7 Distribuição dos clientes por situação profissional 53
5.8 Distribuição dos clientes por estado civil 54
5.9 Distribuição por número de dependentes 55
5.10 Distribuição dos clientes por número de relações familiares 55
5.11 Distribuição da percentagem de depósitos por canal 56
5.12 Número de levantamentos (por mês) 64
5.13 Número de idas à sucursal (por mês) 65
5.14 *Logins* site (por mês) 65
5.15 Contas ativas 65
5.16 NPS score 66
5.17 Distribuição dos cartões *Visa* e *Mastercard* dos cartões *American Express* . . . 90
5.18 Dias necessários à resolução da reclamação 90
5.19 Distribuição dos valores reclamados(em euros/mês) 90
5.20 Erro médio absoluto obtido para cada um dos grupos em função da inclusão ou exclusão da rentabilidade do ano anterior como variável de *input* 105
5.21 Erro médio absoluto (erro de teste) em função do número de folhas e parâmetro de complexidade (cp) 106
5.22 Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp) 107
5.23 Erro médio absoluto (erro de teste) em função do número de folhas e parâmetro de complexidade (cp) 108
5.24 Resultados obtidos pelas Cadeias de Markov. 111

5.25 Erro de treino e teste obtidos com a função de inicialização dos pesos e bias . 112
5.26 Variação do erro de treino e teste em função da *Learning fuction* 112
5.27 Variação do erro de treino e teste em função do *learning rate* 113
5.28 Variação do erro de treino e teste em função do número de épocas 113
5.29 Variação do erro de treino e teste em função do *size* 113
5.30 Variação do erro de treino e teste de acordo com a função de *output* utilizada 113
5.31 Parametrização da MLP 114
5.32 Caracterização dos *clusters* (médias mensais, por variável) obtidos através do método *K-means* 119

A.1 Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp) 129
A.2 Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp) 130
A.3 Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp) 131

Glossário

Accuracy	Métrica usada para avaliar a capacidade preditiva dos modelos de *Machine Learning*.
AHP	Método hierárquico usado para determinar o peso das variáveis. A escolha dos pesos a atribuir a cada variável é feita com base nos conhecimentos (e opiniões) dos especialistas do negócio.
Algoritmos estocásticos	Algoritmos utilizados em análises de tendências, geração de padrões e previsões.
Algoritmos genéticos	São algoritmos meta heurísticos usados para agrupar dados.
Análise preditiva	Combina a descoberta com a análise totalmente automatizada para permitir que o decisor aprenda de acordo com o passado. Usa a modelagem estatística para identificar tendências e padrões, a fim de tomar decisões mais informadas.
Análise de clusters	Partição dos dados em subgrupos. Dentro de cada grupo pretende-se ter uma grande homogeneidade e entre grupos distintos pretende-se que haja heterogeneidade.
Árvores de Decisão C5	Corresponde ao algoritmo sucessor das árvores de decisão C4.5. Comparativamente com o C4.5 caracteriza-se por ser um algoritmo mais rápido que usa menos memória. Além disso, incorpora novas funcionalidades, como, por exemplo, a "variável de classificação de custo".
Churn rate	Taxa de clientes que anualmente deixam de assinar um serviço ou encerram um relacionamento comercial.
CLV de *coorte*	Conjunto de pessoas que registaram o mesmo evento no mesmo período temporal.
CRM	Tecnologias e processos utilizados pelas empresas com o intuito de entender o comportamento dos clientes a fim de melhorar as taxas de aquisição e retenção.
Cross selling	Estratégia de venda de produtos de categorias diferentes daqueles que o cliente já possui. Serve para aumentar a confiança do cliente e diminuir a probabilidade de *churn*.

Customer Lifetime Value	Conceito de gestão de clientes, definido há mais de 30 anos por Kother como "o fluxo de lucro futuro esperado num determinado horizonte temporal de transações com o cliente".
Customer attrition	Ocorre quando os clientes terminam o seu relacionamento com uma determinada empresa. Também designado por *customer churn*, *turnover* ou *defection*.
Discount rate	Converte o valor das receitas futuras para aquilo que são os fatores económicos atuais (como, por exemplo a inflação). Considere-se, a título ilustrativo o seguinte exemplo: se se colocarem 100 euros numa conta bancária com um juro associado de 10% em 12 meses nessa conta existirão 110 euros. Assim, os 110 existentes daqui a 1 ano serão equivalentes aos 100 euros de hoje.
Feature	Variável usada como *input* num modelo de *Machine Learning*.
Feature engineering	Processo que usa o domínio do conhecimento para criar novas *features* a partir dos dados em bruto. Caso o processo de *feature engineering* seja feito de forma correta o poder preditivo dos algoritmos de *machine learning* aumentará.
Gradiente Descendente	É um algoritmo de otimização usado para minimizar algumas funções, movendo-se iterativamente na direção da descida mais íngreme. Em *machine learning*, usa-se para atualizar os parâmetros do nosso modelo.
Homofilia	Tendência para pessoas se relacionarem com pessoas parecidas consigo.
K-means	É um método de *clustering* em que as observações são divididas em K grupos.
Lost-for-good	Cliente deixa de comprar a um dado fornecedor e passa a comprar a outro. Nessas situações, os clientes são difíceis de recuperar.
Maximal relevance	Seleciona as *features* de maior relevância para o *target*. A relevância é caracterizada em termos de correlação/informação mútua usada para definir a dependência entre variáveis.

Mean Decrease Accuracy	É uma maneira de medir a importância da variável. Este indicador permite classificar as variáveis de acordo com as suas habilidades discriminatórias.
mRMR	É um método de *feature selection* que tem como objetivo maximizar a relevância através da seleção das *features* que tem maior relevância para o *target*, sendo a relevância caracterizada em termos de correlação entre variáveis. Este método pretende simultaneamente minimizar a redundância.
Multi Layer Perceptron	Rede neuronal.
Noise sensitive	Representa a sensibilidade do *workflow* ao ruído existente no *dataset*.
Out-of-bag (OOB) error	É um método usado para medir o erro de predição nas *random forest*, *boosted decision trees* e outros modelos de *machine learning*. Este método utiliza agregações de *bootstrap* (*bagging*) para criar sub-conjuntos através do conjunto de treino. O OOB corresponde á média do erro de previsão em cada amostra de treino, utilizando apenas as árvores que não contem essa amostra na sua amostra de *bootstrap*. Assim, a sub-amostragem permite definir uma estimativa *out-of-bag*, que contribui para a melhoria de desempenho de predição, avaliando predições sobre as observações que não foram usadas na construção do base do próximo *learner*.
Overfitting	Termo usado para descrever o ajuste excessivo do modelo aos dados de treino. Quando isto acontece o modelo perde a capacidade de generalização, mostrando-se por isso ineficaz na previsão de resultados.
Programas de referência	Técnica usada para adquirir clientes: a empresa compensa os clientes que trouxerem novos clientes para a empresa. No entanto, não há evidências de que os clientes adquiridos por esses programas sejam mais valiosos do que os clientes adquiridos de outra maneira. Os clientes referidos tem uma margem de contribuição alta embora essa diferença seja desgastada com o tempo.

Propensão ao produto	Probabilidade do cliente comprar os produtos "recomendados" pela empresa.
Random Forest	São métodos *ensemble* que tanto podem ser usados para problemas de classificação como de regressão. Operam construindo uma infinidade de árvores de decisão no momento do treino e produzindo a classe (classificação) ou previsão média (regressão) das árvores individuais. Tem a vantagem de não ser propensas ao *overfitting*.
Rede Neural de Kohonen	Algoritmo com capacidade de organizar dados complexos em *clusters* tendo em consideração as suas relações entre eles. Este método é ideal para problemas onde os padrões são desconhecidos ou indeterminados.
Redes sociais-SNA	Mapeamento de um conjunto de entidades (nós) e medição das relações existentes entre essas entidades. A rede resulta do fluxo de valores que ocorre entre os nós. Salienta-se que os nós podem ser grupos ou organizações enquanto que arestas representam relacionamentos entre nós. As técnicas de SNA podem ajudar a determinar nós influentes dentro de uma determinada rede.
Regra 80/20	80% dos lucros de uma empresa são produzidos por 20 % dos seus clientes. Os restantes clientes (80%) produzem o equivalente a 20% dos lucros da empresa.
SOM	Rede neural treinada usando aprendizagem não supervisionada para produzir uma representação discreta de baixa dimensão (geralmente bidimensional) das variáveis de *input*.
Up selling	Ação de vender a determinado cliente atualizações ou outros complementos de produtos adquiridos anteriormente.
Valor potencial do cliente	Lucros que se espera obter de determinado cliente quando este usa os serviços adicionais.
Value Network	Reflete o valor de intercâmbio dinâmico, tangível e intangível numa rede de clientes que negoceiam entre si.

Siglas

AID	*Automatic Interaction Detection*
ANN	*Artificial Neural Network*
CART	*Classification and Regression Trees*
CE	*Customer Equity*
CLV	*Customer Lifetime Value*
cp	*complexity parameter*
DTMC	*Discrete time Markov chains*
GLM	Generalized Linear Model
MAE	*Mean Absolute Error*
MLP	*Multi Layer Perceptron*
NBD	Distribuição Binomial Negativa
RF	*Random Forest*
SMO	*Sequential Minimal Optimization*
SNA	*Social Network Analysis*
SOM	Mapa Auto-Organizável
SVM	*Support Vector Machines*

Capítulo 1

Introdução

Neste capítulo é apresentado o conceito de Customer Lifetime Value e as motivações para o estudo deste tema. É feita também uma contextualização do setor bancário (dado que o dataset utilizado provêm de uma instituição bancária portuguesa). Quanto á estrutura, esta dissertação encontra-se organizada da seguinte forma:

***Capítulo 1**: Contextualizar o tema e apresentar as motivações e objetivos deste estudo;*

***Capítulo 2**: Revisão da literatura;*

***Capítulo 3**: Descrever e caracterizar o dataset que será utilizado ao longo deste estudo;*

***Capítulo 4**: Explicar a teórica subjacente aos métodos utilizados para modelar o target;*

***Capítulo 5**: Apresentar os resultados;*

***Capítulo 6**: Discutir de forma crítica dos resultados obtidos;*

***Capítulo 7**: Referências bibliográficas.*

Salienta-se que todos os conceitos técnicos encontram-se definidos no glossário que se encontra nas páginas iniciais desta dissertação.

1.1 Definição

O "Valor de tempo de vida do cliente" (da literatura anglo-saxónica, *Customer Lifetime Value*, CLV) é um conceito usado para medir o valor futuro de cada cliente, baseando-se: (1) no valor de contribuição económica direta, (2) em fatores relacionados com a sua contribuição económica direta e (3) na amplitude do valor de contribuição económica

indireta, contabilizando outros fatores como a volatilidade e vulnerabilidade dos fluxos de caixa dos clientes [30].
Nos últimos anos, as estratégia de *marketing* das empresas tem dado grande importância à manutenção do relacionamento contínuo com os seus clientes. Nesse sentido, o CLV tem sido uma métrica usada nas abordagens de *marketing* de relacionamento, pois permite inferir sobre a potencial receita futura gerada por cada cliente. Por outras palavras, "o valor da vida útil do cliente" pode-se definir como "o lucro líquido total que uma empresa pode esperar de um cliente" durante o período em que existe qualquer relacionamento com a empresa.
O CLV pode ser considerado uma métrica desagregada ou agregada. Quando se calcula o CLV enquanto métrica desagregada o objectivo é encontrar clientes com elevada rentabilidade futura estimada. Tal permite alocar recursos para evitar que estes abandonem a empresa [45]. Além disso, o CLV de clientes atuais e futuros também é uma boa medida de valor global de uma empresa [11].
Uma das principais dificuldade do cálculo do CLV é o facto dos clientes apresentarem padrões comportamentais distintos e difíceis de modelar: há conjuntos de ações que não são realizadas com uma periodicidade associada, como por exemplo a aquisição de crédito. Na figura 1.1 está representada, a título ilustrativo, a frequência de compras para um determinado conjunto de clientes. Para cada cliente representa-se com um "traço" o momento em que adquiriu produtos da empresa considerada. Com o histórico de compras, representado do lado esquerdo da imagem, pretende-se determinar a frequência e distribuição dos padrões de compra futuros para cada um dos clientes.O primeiro cliente adquiriu produtos com muita frequência no início do período considerado. No entanto, a partir de determinado momento, deixa de fazer aquisições. Por outro lado, o segundo cliente fez aquisições ao longo de todo o período, mas com intervalos de tempo mais espaçados. Dado isto, e analisando apenas uma variável de forma simplificada a questão que se coloca é: "Qual dos dois clientes apresenta um valor mais elevado para a empresa?"

Dado isto e segundo Kumar & George (2007) um modelo básico para o cálculo do CLV de um dado cliente, i, num determinado intervalo de tempo, t, resulta da diferença entre a receita gerada, R, e os custos, C, que estão associados à geração da receita, obtendo-se desta forma o lucro líquido. O *discout rate*, δ, está associado à "transformação", do valor de lucro líquido em lucro atual. Para calcular o CLV podem ser considerados T períodos, obtendo-se [31],

$$CLV_i = \sum_{t=1}^{T} \frac{R_t - C_t}{(1+\delta)^t}. \tag{1.1}$$

1.2 Objetivos

O principal objetivo deste estudo é a modelação do CLV com o intuito de construir modelos preditivos que permitam estimar o CLV futuro de cada cliente. Outro dos objetivos

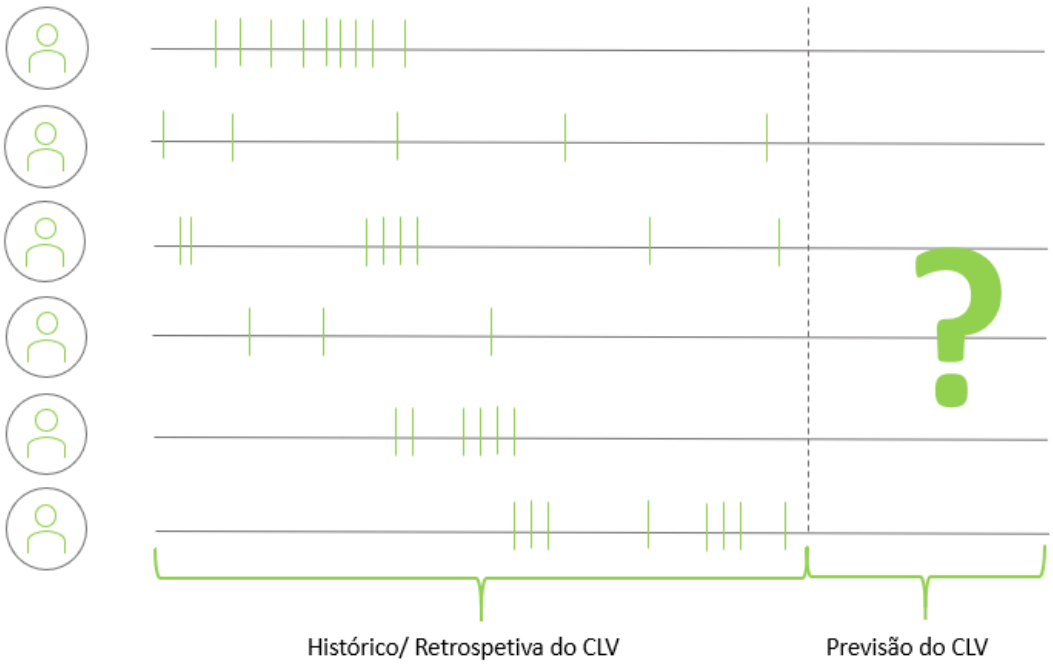


Figura 1.1: Distribuição de padrões de compra por cliente [57].

desta dissertação é obter conhecimentos aprofundados sobre os clientes de modo a sugerir o desenvolvimento de campanhas personalizadas e rentáveis.
Com esse propósito, definem-se como objetivos específicos a implementação de modelos *Machine Learning* e *Deep Learning* e avaliação da sua performance. Define-se também como objetivo específico desta dissertação a comparação entre os vários métodos implementados e a realização de uma análise de *clusters* de modo a caracterizar e sugerir abordagens de *marketing* estratégicas para cada *cluster*.
Na figura 1.2, apresenta-se a *timeline* seguido na realização da dissertação de mestrado.

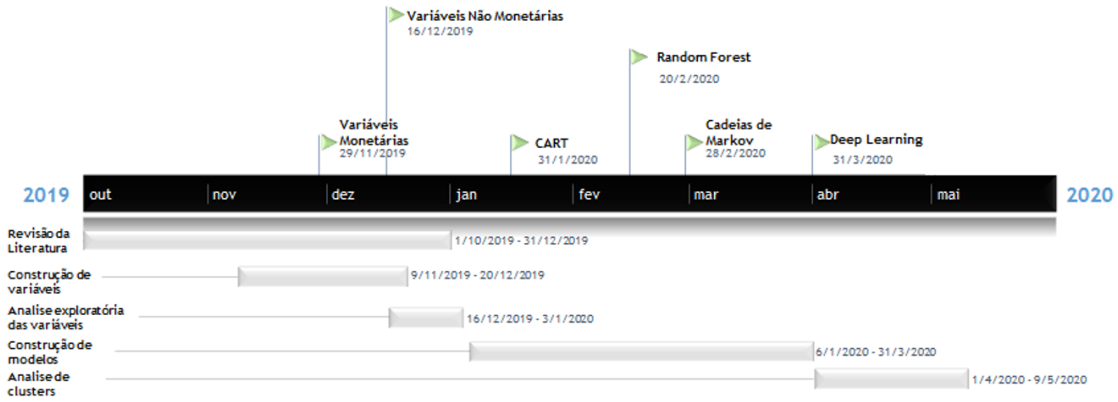


Figura 1.2: *Timeline* da dissertação

1.3 Motivação

Atualmente, as entidades bancárias enfrentam um momento de grande competitividade devido ao crescimento e aparecimento de novos bancos e à facilidade de mudança de entidade bancária, que hoje em dia, já é possível fazer *online* em poucos minutos. Além disso, os clientes tornaram-se mais exigentes, mudando de banco caso os benefícios e facilidades que desejam não sejam concedidos pelo banco em que se encontram. Face a este cenário é importante para as empresas bancárias o cálculo de métricas, como o CLV, que lhes permitem estimar a rentabilidade futura de cada um dos seus clientes (expressa em unidades monetárias). Tendo em conta esse valor, as empresas podem planear os investimentos nos seus clientes (como, por exemplo em campanhas de *marketing* ou nas taxas de descontos aplicadas na compra de produtos) de uma maneira mais ponderada.
Salienta-se, no entanto, que a utilização deste conceito exige um estudo aprofundado do cliente, pois é necessária uma análise cuidada de todas as variáveis que afetam direta e indiretamente a rentabilidade futura do cliente. Esse conhecimento do cliente poderá ser posteriormente utilizado para melhorar a relação com os clientes e diminuir o *customer attrition*.
Em termos gerais, podem-se definir como principais vantagens da utilização do CLV os seguintes aspetos:

- Identificar a rentabilidade dos clientes;
- Determinar os investimentos para retenção dos clientes;
- Desenvolver ações de *marketing* rentáveis;
- Implementar estratégias de *marketing* alternativas;
- Prever a prevalência e a satisfação do cliente;
- Calcular a posição da empresa e do seu valor no mercado;
- Desenvolver um suporte para uma gestão do cliente a longo prazo;
- Ajustar a estratégia empresarial tendo em conta o estilo de vida do cliente;
- Escalar o nível de dados com a tecnologia disponível;
- Desenvolver uma abordagem centrada numa gestão rentável do cliente;
- Determinar os parâmetros/variáveis com maior contributo para o lucro de cada segmento de clientes, facilitando assim a tomada de decisão;
- Destacar o valor de diferentes ativos (bens adquiridos que geram retorno financeiro) para a empresa. Os ativos não relacionais são valiosos se conseguirem aumentar o valor dos ativos do cliente;

- Estimar o efeito de várias atividades organizacionais no cliente;
- Desenvolver valores e produtos baseados na segmentação comportamental. Evidencia-se que a segmentação de clientes de alto risco apresenta menor precisão, pois estes clientes apresentam propensão para mudar o seu comportamento caso, estejam expostos a incentivos aliciantes;
- Identificar clientes com propensão ao *churn* de modo a poderem ser inseridos em programas de "recuperação" e evitar que abandonem a empresa;
- Identificar os fatores que provocam *churn*.

Outro aspeto que tem grande importância para o sucesso de uma empresa é a sua capacidade de atrair novos clientes, mantendo ou aumentando as vendas com os clientes existentes. Salienta-se que clientes mais satisfeitos tem maior probabilidade de comprar e recomendar os produtos continuando a gerar lucro a longo prazo. Os pequenos aumentos nas taxas de retenção de clientes tem grande impacto no lucro da empresa, bem como, na aquisição de vantagem competitiva em relação aos seus pares [2].

1.4 Contextualização

Os dados usados para a realização deste estudo provêm de uma entidade bancária. Neste caso particular, e de acordo com Khajvand & Tarokh (2011) e Haenlein *et al.* (2007), os modelos de CLV devem satisfazer as seguintes condições [22, 29]:

- Incluir as transações discretas que podem ocorrer apenas uma/escassas vezes durante o período em que o cliente mantém contacto com a empresa (por exemplo, operações de manutenção de conta);
- Serem fáceis de interpretar de modo a garantir a sua aplicabilidade em contextos comerciais variados.

Este interesse e necessidade por parte das instituições financeiras em desenvolver métricas que lhes permitam uma maior rentabilidade surgiu, em parte, devido às consequências da crise financeira que 2008 afetou todo o mundo, e ao aumento da competição *cross-border* no retalho bancário com a introdução do euro, que impactou negativamente as margens e rentabilidade das organizações financeiras. Para solucionar esse problema, os bancos começaram a fazer uma comercialização dos produtos orientada para as necessidades dinâmicas do cliente e que permitissem uma medição mais precisa do retorno do investimento em *marketing* [35]. Além dos concorrentes diretos, devido ao avanço tecnológico que se verificou nas últimas décadas, apareceram também "concorrentes" inovadores, como a: *Apple*, *Google* ou algumas *start-up* que obrigaram o retalho bancário a renovar-se e a desenvolver mecanismos que lhes permitissem ter uma melhor relação com os clientes

(por exemplo, o desenvolvimento de *apps* que permitem a aquisição de produtos financeiros de uma forma rápida e simples). Para manter a vantagem competitiva, outrora mais evidenciada, apostou-se na manutenção dos pontos de diferenciação, inovando através da personalização de serviços e produtos, bem como na oferta de produtos adequados às necessidades do cliente naquele momento específico. Esta abordagem baseia-se na análise do comportamento e preferências passadas [20].
O ciclo de vida de um relacionamento comercial inclui 3 fases [49]:

1. Identificação: Tem como objetivo identificar indivíduos rentáveis e com probabilidade de se tornarem cliente da organização. Para "encontrar" estes indivíduos recorre-se a técnicas de *clustering* e segmentação para explorar dados históricos e pessoais.

2. Retenção: Conjunto de ações desenvolvidas com o intuito de garantir a fidelidade dos clientes e reduzir o *churn*.

3. Desenvolvimento: Conjunto de ações desenvolvidas com o intuito de aumentar a quantidade de transações e de compra de produtos de modo a maximizar a rentabilidade do cliente.

O *Cross-selling*, *up-selling* e a aquisição ou retenção do cliente (dependendo dos casos) são indicadas como sendo as principais atividades para aumentar o valor do cliente. A aquisição apresenta em média, custos 5 a 20 vezes superiores aos custos associados à retenção porque a falta de informação sobre os novos clientes dificulta o enfoque nos produtos que o cliente poderá ter interesse em adquirir [25]. A posse de produtos *cross-selling* ou/e produtos *up-selling* também podem ser fatores a ter em conta no cálculo do CLV, pois estes são fatores que permitem inferir sobre o grau de satisfação e fidelização do cliente bem como sobre a relação que o cliente eventualmente terá com outras empresas concorrentes. Salienta-se que clientes que possuem produtos *cross selling* apresentam em regra menor probabilidade de mudar de fornecedor.
Convém ainda realçar que a retenção e a capitalização são afetados por fatores não observáveis com as forças macroeconómicas [2].

Capítulo 2

Estado da arte

Neste capítulo são apresentadas as várias abordagens que diversos autores adotaram para o cálculo e modelação da variável target, o CLV.

2.1 Cálculo do *CLV*

Genericamente, existem duas abordagens para o cálculo do CLV:

- **Abordagem agregada**: tem como objetivo o cálculo do valor médio do CLV de *coorte*. Esta medida designa-se por *Customer Equity* (CE);
- **Abordagem individual**: o CLV é calculado para cada cliente (individualmente).

Kumar & George (2007) fizeram uma discussão detalhada de comparação das abordagens acima descritas, onde concluíram que as abordagens agregadas apresentavam piores resultados. Assim, ao longo deste estudo, o CLV será abordado numa perspetiva individual [31].

2.1.1 *Twelve Month-Prune*

Dada a complexidade inerente ao comportamento humano o CLV tem vindo a ser aperfeiçoado ao longo dos anos de modo a conseguirem-se fazer previsões mais precisas do valor do cliente para a empresa. Antes de 1960, os comerciantes, de modo a saberem quais eram os seus clientes mais fiéis, e que por isso pressupunham que seriam os mais lucrativos, criaram as listas *12 month-prune*. Nestas listas constava o nome de todos os clientes que tinham feito alguma compra no último ano. Assim, se cliente não tivesse comprado nada no último ano, era retirado da lista. Este método foi usado inicialmente por empresas

americanas que tinham clientes rurais [2].

2.1.2 RFM

Do aperfeiçoamento do método ***12 Mounth Prune*** (secção 2.1.1), em que a única *feature* considerada para o cálculo do CLV era a recência (data em que o cliente efetuou a última compra), surgiu o método RFM, que além da recência (*R* - recency), tinha também em conta para o cálculo do CLV, a frequência com que o cliente faz compras (*F* - frequency) e o valor monetário das compras (*M* - *monetary*). Este é um dos métodos mais utilizados para o cálculo do CLV sendo que, nos últimos anos tem-se optado por usar as variáveis R, F e M como *inputs* para modelos de *Machine Learning*, ou outros, como, por exemplo, séries temporais.
Khajvand &Tarokh (2011), e Hasheminejad & Khorrami (2018) utilizaram séries temporais para fazer a projeção do valor futuro do cliente, sendo que, em ambos os artigos, as variáveis R, F, M foram usadas como *inputs* [24, 29]. Começou-se por utilizar algoritmos de *clustering* para segmentar os clientes. Seguiu-se o cálculo do valor de cada segmento foi feito com base no modelo RFM ponderado, em que o peso associado a cada variável foi determinado pelo método AHP (técnica em que os decisores atribuem diferentes pesos às variáveis de *input*). A atribuição de pesos é feita com base num critério fixo e por comparação com os pares de opção fornecidos. O critério que apresenta maior importância para a tomada de decisão terá uma maior pontuação. De seguida utilizou-se o modelo ARIMA (*Autoregressive Integrated Moving Average*) para prever a rentabilidade de cada um dos segmentos.
Dada a complexidade do comportamento humano vários autores consideram importante ter em conta outras variáveis na predição do valor futuro do cliente. Noori (2015) incluiu no seu trabalho uma variável (denotada por D) referente à classificação do depósito do cliente, surgindo assim o RFMD [39]. Yoseph & Heikkila (2019) consideram importante incluir uma variável, C, que quantificasse a mudança de comportamento de compra bem como o sentido da mesma [61]. As variáveis demográficas também são, muitas vezes, tidas em conta para a modelação do CLV como foi proposto por Rabiei (2015) que começou por fazer a extração das variáveis RFM ponderadas, com recurso AHP, para cada cliente tendo de seguida procedido à segmentação com recurso ao método*K-means* tendo em consideração as variáveis demográficas [43]. De seguida calculou-se o CLV para cada grupo de clientes. Neste caso, pretendia-se ter como *output* a probabilidade de resposta a determinada campanha ou oferta de produto. Neste modelo de classificação os clientes são posteriormente agrupados em duas classes: respondentes e não respondentes. O algoritmo utilizado foi uma árvore de decisão C5. Estes algoritmos têm a vantagem de conseguir gerar regras que podem ser traduzidas por linguagem natural.
Nos casos em que as variáveis de *input* fornecidas a determinado modelo tem pesos/importâncias diferentes (podendo estes ser decididas, por exemplo, com base no método

AHP) diz-se que se está a utilizar o WRFM (em que o W representa a inclusão do peso atribuído a cada uma das variáveis). Depois de numa fase inicial Ayoubi (2016) ter aplicado o WRFM, de seguida utilizou a rede neural de *Kohonen* para organizar dados em *clusters* tendo em conta as suas relações [3]. Como este algoritmo consegue reduzir a dimensionalidade e manter a representação real, após a sua utilização torna-se mais fácil a identificação dos *clusters* mais valiosos. Este algoritmo é considerado um Mapa Auto-Organizável (SOM).

Após o processamento inicial dos dados recorrendo ao método AHP, e uso do método *K-means* e análise discriminante, quer Hajipour & Esfahani (2019) quer Haenlein *et al.* (2007) optaram por usar Cadeias de Markov e modelos CART (classificação de árvores de regressão) [22, 23]. Este tipo de modelação tem a particularidade de conseguir bons resultados quer para fluxos de transações discretos, quer fluxos de receitas contínuos. Permite também usar variáveis em diferentes escalas, sem se correr o risco de haver inconsistências bem como a capacidade de análise dos clientes através da criação de grupos homogéneos.

Num processo de Markov (processo estocástico no qual a probabilidade de transição entre 2 estados discretos depende apenas das propriedades do estado imediatamente anterior) cada um dos grupos homogéneos é considerado um estado da natureza entre os quais é permitido que os clientes transitem através de um processo de Markov, de primeira ordem. Da combinação de diferentes processos de Markov resulta uma cadeia de Markov, com as respetivas probabilidades de transição, que se encontram resumidas numa matriz de transição. Para estimar a probabilidade de transição determinou-se, com recurso a uma CART, o estado de natureza a que cada cliente pertencia no início e final do intervalo de tempo considerado. Pela razão entre o número de clientes que transitaram de estado e o número total de clientes estimou-se a frequência de transição, que serviu como *proxy* da probabilidade de transição adjacente. Salienta-se que Haenlein *et al.* (2007) optou por realizar separadamente análises das CART para diferentes faixas etárias [22]. Esta metodologia apresenta como desvantagem o facto de assentar no pressuposto de que a matriz de transição será estável e constante ao longo do tempo, o que parece apropriado para previsões de médio prazo, desde que não haja razões para mudança drástica no comportamento de consumo. No entanto, este pressuposto não é uma suposição aceitável para previsões de longo prazo.

2.1.3 Métodos determinísticos

Neste subcapítulo serão apresentadas várias fórmulas de cálculo do CLV. Salienta-se que não existe uma maneira apropriada para calcular o tempo de vida do cliente. A natureza do negócio e do objetivo da modelação tem de ser tidos em conta na escolha da fórmula de cálculo [12].

Entre as fórmulas revistas para calcular o CLV no nível individual, uma das mais simples e mais utilizada foi proposta por Jain & Singh (2002) [27]. Salienta-se, que com o intuito

de conseguir uma melhor adequação desta fórmula ao ramo de negócio considerado no estudo, esta têm sofrido algumas alterações no denominador por parte de vários autores, como, por exemplo Kumar & George (2007), visto que, o denominador é responsável pela conversão do valor líquido em valor atual, e essa taxa de conversão varia de negócio para negócio. O numerador representa o lucro líquido obtido em cada período [26].
A simplicidade dessa fórmula é considerada por vários autores como muito atraente e, embora não tenha parâmetros específicos para representar interações, como venda cruzada, elas podem ser incluídas nas receitas de períodos futuros. Outra das críticas apresentadas a esta fórmula é facto de não considerar custos indiretos, como, por exemplo os custos de *marketing*. Mesmo assim, este modelo de cálculo do CLV é considerada adequado para apoiar indiretamente ações da empresa, como aquisição, retenção, venda cruzada de clientes, entre outras [46].
Assim, e segundo Jain & Singh (2002), o CLV foi definida da seguinte forma [1]:

$$CLV_i = \sum_{t=1}^{T} \frac{(R_t - C_t)}{(1+\delta)^{t-0.5}}. \tag{2.1}$$

Nesta dissertação como *target*, ou seja, CLV, utilizou-se a Rentabilidade Líquida, cuja fórmula de cálculo vai ao encontro da fórmula 2.1 apresentada por Jain & Singh (2002). A Rentabilidade Líquida considera como receita as margens de recurso, as margens a crédito e as comissões. Como custos considera a imparidade e os custos de transação. Para definir o *discount rate* utilizou-se a Euribor, por ser considerada a "taxa de referência mais importante nos mercados financeiros europeus" [17]. Num caso em que um cliente apresente um CLV de 100 euros isso significa que, após serem subtraídas às receitas geradas pelo cliente, os custos que as suas operações e imparidades tiveram para o banco, nos últimos 12 meses, o banco lucrou 100 euros com esse cliente. No caso de um cliente que apresente um CLV de -100 euros significa que os gastos que o banco teve com esse cliente foram superiores aos lucros que ele/a gerou para o banco e, por isso, ao fim de 12 meses, o banco perdeu 100 euros com esse cliente.
A escolha desta fórmula para calcular o CLV deveu-se ao facto de esta apresentar simplicidade no cálculo. Outra das razões para usar esta fórmula é o facto de esta apresentar um bom ajuste ao negócio bancário, tendo por isso sido anteriormente utilizada noutros estudos desta área [27].
Em 2004, Gupta, Lehman e Stuart [11] propuseram uma atualização para a fórmula anterior: os custos de aquisição, AC, passaram a ser considerados no cálculo do CLV. De modo a avaliar a retenção, esta nova fórmula também inclui a probabilidade de o cliente ainda se encontrar ativo no período considerado, P(Ativo). Assim, o CLV passou a ser definido por [1]:

$$CLV_i = \sum_{t=1}^{T} \Big[\frac{(R_t - C_t) \times P(Ativo)_{i,t}}{(1+\delta)^t} \Big] - AC. \tag{2.2}$$

[1] Salienta-se que os parâmetros que constituem a fórmula de cálculo, já foram definidos anteriormente.

A razão pela qual se decidiu excluir esta fórmula deve-se ao facto de em cenários de negócio não contratuais, como é o caso de retalho bancário, ser difícil medir com exatidão a probabilidade de o cliente ainda se encontrar ativo.

A *Customer equity* corresponde ao valor patrimonial do cliente e pode ser medidos quer usando abordagens agregadas, quer usando abordagens desagregadas.

Blattberg & Deighton (1996) desenvolveram uma abordagem de nível agregado usando *customer equity* para balancear a aquisição e retenção [7].

Berger & Nasr (1998) propuseram que o conceito de CLV assentasse nos seguintes pressupostos: 1) As vendas ocorrem uma vez por ano; 2) Os gastos com a retenção, M [2], a taxa de retenção, r, e a margem bruta de contribuição anual, GC, são calculadas anualmente e permanecem constantes ao longo do tempo. Estes autores afirmaram também que existia uma diferença significativa entre CE e CLV, segundo eles, o CE deveria ter em atenção os custos de aquisição do cliente [5, 32]. Sob essas premissas, o CLV é calculado da seguinte forma [1]:

$$CLV = GC \times \sum_{t=1}^{T} \left[r^t / (1+\delta)^t \right] - M \times \sum_{t=1}^{T} \left[r^{t-1} / (1+\delta)^{t-0.5} \right]. \tag{2.3}$$

Apesar de a retenção ser um dos parâmetros usados de forma recorrente no cálculo do CLV existem diferenças conceituais em termos de contabilização de clientes existentes, perspetivas de aquisições e período de projeção existem nas diferentes abordagens [32].

Esta fórmula de cálculo do CLV não fazia sentido ser utilizada nesta dissertação uma vez que se pretende que cada cliente tenha um valor de CLV, e isso só possível se se optar por abordagens desagregadas.

A abordagem de Rust *et al.* (2000) sugere que o cálculo do CLV seja feito utilizando informações sobre a marca focal e as marcas concorrentes para modelar a aquisição e retenção de clientes no contexto de troca de marca. Para uma determinada marca, *j*, o cálculo do CLV dos seus clientes envolve entrevistas nas quais os clientes respondem a perguntas relacionadas com o número de compras que efetuam, *f*. De seguida calcula-se a probabilidade de o cliente voltar a comprar produtos daquela marca, definida por B, recorrendo para isso ao cálculo da matriz de transição de Markov. Neste modelo de cálculo também é considerado o volume de compras expectável por parte de cada cliente, V, bem como a margem de contribuição expectável, por unidade da marca comprada, definida por π [48]. O valor da vida útil, CLV_{ij} do cliente i para a marca j, é dado por [1]:

$$CLV_{ij} = \sum_{t=0}^{T_{ij}} \frac{1}{(1+\delta_j)^{t/f_i}} V_{ijt} \times \pi_{ijt} \times B_{ijt}\,. \tag{2.4}$$

Na abordagem de Rust *et al.* (2000) faz-se o cálculo do CLV considerando diferentes marcas do produto em causa [48]. Como neste estudo apenas são considerados os dados de uma instituição financeira não faria sentido considerar esta fórmula. Stahl (2003)

[2] As despesas promocionais, são um exemplo de gastos com a retenção.

propôs um método para analisar a relação entre o valor de tempo de vida do cliente e o valor do *shareholder* [55]. Wu *et al.* (2005) consideraram que a influência social, o valor potencial do cliente e o valor de *network* [60] deveriam ser parâmetros a ter em conta no cálculo do CLV. Neste caso, o CLV, é composto por [55]:

- *Base Potencial*, Bs: Corresponde ao *cash flow* gerado na compra de produtos e serviços. Este parâmetro é um bom indicador da quantia de dinheiro que será gasta no futuro. Para fazer essa previsão, os autores recorreram a *times series*, conseguindo assim mesurar as "flutuações" e regularidades dos gastos.
- Valor potencial, P: *Cash flow* resultante de *cross-selling*, *uptrading*, e aumentando *"share of the wallet"*;
- *Valor de Networking*, N: *Cash flow* resultante da publicidade "boca a boca";
- *Potencial Learning*: *Cash flow* resultante do conhecimento do cliente através da relação que foi criada.

Segundo Wu *et al.* (2005) o CLV seria dado por [1]:

$$CLV_i = \sum_{t=1}^{T} \frac{\left(Bs_i(t) + P_i(t) + N_i(t)\right) \times r_i(t) - AC_i(t)}{(1+\delta)^t}. \tag{2.5}$$

A principal crítica apresentada a esta abordagem é o facto de a proposta de Stahl não apresentar nenhuma solução para medir o contributo dos "programas de referência"[55, 60]. No dados fornecidos para a elaboração desta dissertação não havia informação que permitisse medir a influência social. Assim, tornou-se inviável a utilização desta fórmula.

Venkatesan & Kumar (2004) usaram a abordagem *always-a-share* (modelo que assume que os clientes podem facilmente começar a comprar parte ou a totalidade dos produtos que necessitam a outra empresa) na sua proposta para o cálculo do CLV, para isso, optaram por calcular a soma cumulativa de *cash flow* com desconto dos WACC *(Weighted Average Cost of Capital)* do cliente ao longo do seu tempo de vida com a empresa. Os autores sugerem também a incorporação dos custos de marketing, c, do cliente i no canal m do período t. Outra das novidades desta fórmula de cálculo do CLV é a incorporação do número de contactos feitos ao cliente, representados na equação por x. Assim, e segundo estes autores, o CLV seria dado por [1]:

$$CLV_i = \sum_{t=1}^{T_i} \frac{GC_{i,t}}{(1+\delta)^{t/f_i}} - \sum_{t=1}^{T} \frac{\sum_m c_{i,m,t} \times x_{i,m,t}}{(1+\delta)^{t-1}}. \tag{2.6}$$

Neste caso as principais componentes envolvidos no cálculo do CLV são a frequência de compra, as margens de contribuição e os custos de *marketing*.
Em negócios não contratuais torna-se difícil saber quando o cliente deixa de estar ativo, pois neste cenário não existe uma subscrição periódica. Dada essa necessidade Fader *et al.* (2007) decidiram propor uma fórmula de cálculo do valor esperado do CLV, E(CLV), que

será calculado tendo em conta o valor esperado de *cash flow*, $E[v(t)]$ e a probabilidade do cliente ainda se encontrar ativo, que neste caso foi calculada com recurso ao modelo de Distribuição Binomial Negativa (NBD) [18]. Crowder *et al.* (2007) na sua abordagem para o cálculo do CLV considera para cada cliente a taxa de acumulação de lucro *v(t)*, onde *t* representa o tempo. O tempo que o cliente se mantém na empresa representado por *T*, e chamado de *tenure*, é uma variável aleatória com função de distribuição F(t) [12]. Dado isto, os autores propuseram a seguinte fórmula de cálculo:

$$E(CLV) = \sum_{t=0}^{\infty} \frac{E[v(t)] \times P(Ativo)}{(1+\delta)^t} \tag{2.7}$$

$$\text{sendo } \; E[v(t)] = E\left[\int_0^t v(s)ds\right] = \int_0^{\infty} v(s)1 - F(s)ds.$$

Berger *et al.* (2006) no modelo de cálculo do CLV que desenvolveram não tinha em consideração a taxa de retenção tal como também não consideravam os custos de aquisição. Assim, o "CLV seria uma função dos lucros líquidos futuros (depois de deduzir custo dos produtos vendidos e outros custos marginais variáveis custos)" [4]. Os custos futuros referem-se àqueles que são cobrados a clientes individuais. Este modelo é traduzido pela seguinte fórmula:

$$CLV_i = \sum_{t=1}^{T} \frac{(FutureGrossProfits_{it} - FutureCosts_{it})}{(1+\delta)^t}. \tag{2.8}$$

Bolton, Lemon e Verhoef (2004) consideraram 3 dimensões que deveriam ser tidas em conta no cálculo do CLV quanto á relação entre os clientes e a empresa: 1) *length*, 2)*depth*, 3) *bradth*. De modo a satisfazer essas condições foram incluídas as seguintes variáveis no cálculo do CLV: $P(retention)_{i,t}$, que representa probabilidade de o cliente *i* continuar a manter uma relação com a empresa no tempo *t* (tempo de duração da relação); $Product_{i,j,t}$, que representa as compras do produto ou serviço *j* pelo cliente *i* no tempo *t*; $Usage_{i,j,t}$ que representa a taxa de uso do produto ou serviço *j* pelo cliente *i*. Neste caso o valor CLV será dado por [8]:

$$CLV_{i,j} = \sum_{t=0}^{T} \frac{P(retention)_{i,t} \times (Product_{i,j,t} \times Usage_{i,j,t} \times GC_{j,t})}{(1+d)^t}. \tag{2.9}$$

2.2 Modelação do *CLV*

Na modelação do CLV tanto podem ser usados modelos estocásticos como modelos determinísticos, no entanto, a escolha do tipo de modelo a utilizar depende se o cenário é contratual ou não contratual [15]. Os modelos determinísticos são usados maioritariamente em cenários contratuais, ou seja, quando é feito um contrato entre a empresa e o cliente e existem prazos para efetuar os pagamentos. Neste cenário é mais fácil modelar o comportamento do cliente. No cenário não contratual, usando-se modelos estocásticos,

pois as transações ocorrem aleatoriamente, e é muitas vezes necessário calcular a probabilidade de um cliente estar ativo e como este se comportará [19].
De seguida serão apresentadas alguns dos métodos que ao longo dos anos foram utilizados na modelação do CLV.

2.2.1 Métodos Estatísticos

Nas décadas de 60 e 70 começaram-se a usar-se técnicas estatísticas para a modelação do CLV. Inicialmente, foram usadas abordagens estatísticas clássicas como a Regressão. Posteriormente, usaram-se outros métodos estatísticos como a *Automatic Interaction Detection* (AID) ou a análise discriminante [2].
Budale & Mane (2013) assumiram que a detecção de fraude, a propensão ao produto e o risco de *churn* são "fatores" que quando considerados em conjunto permitem fazer uma boa previsão do CLV do cliente e tomar decisões de *marketing* assertivas de modo a aumentar o valor do cliente [10]. Para modelar estes autores usaram a classe de modelos designados por Generalized Linear Model (GLM). Nestes modelos assume-se que cada uma das variáveis dependentes é gerada a partir de uma determinada distribuição. Num GLM há sempre 3 componentes: o valor esperado da variável resposta, o preditor linear (que é a combinação linear das variáveis independentes) e a função de ligação entre estes dois.
Neste contexto, autores como Mzoughia *et al.* (2018) recorreram a distribuições conhecidas para modelar o comportamento e/ou transações dos clientes, para de seguida obter o valor do CLV [38]. A utilização da distribuição de *Conway-Maxwell-Poisson*, que utiliza 2 parâmetros, para modelar as transações dos clientes oferece uma maior flexibilidade e ajuste aos dados discretos do mundo real quando comparado com outras distribuições que apenas utilizam 1 parâmetro, como, por exemplo a distribuição de *Poisson*.
Comparando a distribuição de *Pareto/ NBD*, que assume: que o número de transações de um cliente quando este se encontra "ativo", ou seja, realiza transações com determinada frequência segue uma distribuição de Poisson; o tempo entre as transações segue uma distribuição exponencial e taxas de frequências de transações seguem uma distribuição gama (uma vez que nem todos os clientes apresentam necessariamente as mesmas taxas de frequências de transações), com a distribuição de *Conway-Maxwell-Poisson* verifica-se que esta última apresenta uma melhor performance, que é justificada pela existência de erros na modelação na frequência das transações quando se recorre à distribuição de *Poisson*, uma vez que esta assume a equi-dispersão dos dados, mas muitas vezes, na realidade os dados apresentam uma superdispersão ou subdispersão. Dado isto, consegue-se afirmar que segundo o modelo de *Pareto/NBD*, a distribuição de Pareto fornece a distribuição de probabilidade do cliente ainda estar ativo, enquanto que a NBD fornece a distribuição do número de transações durante as fases ativas da vida do cliente. Quanto ao *output*, este modelo, apresenta apenas quatro parâmetros de saída o que se revela claramente

insuficiente para vincular o comportamento de compras individuais a informações sócio-demográficas de modo a conseguir-se prever o comportamento de novos clientes.
A distribuição de *Conway-Maxwell-Poisson* integra 3 distribuições conhecidas (*Bernoulli*, *Poisson*, *Geométrica*), sendo os dados modelados através de 2 parâmetros. Para estimar esses parâmetros, os autores usaram o algoritmo de simulação de Monte Carlo via Cadeias de Markov (MCMC). Outra das conclusões deste estudo foi que o uso do paradigma Bayesiano fornece estimativas individuais dos clientes que ajudam a vincular os comportamentos de compra a características sócio-demográficas. De seguida utilizaram o *K-Means* para agrupar os clientes de modo a criar segmentos. O número ideal de *clusters* foi obtido otimizando o critério de informação Bayesiana. A escolha deste método deveu-se ao facto de este ter capacidade de fornecer informações específicas e individuais para apoiar decisões de *marketing*.

2.2.2 *Machine Learning*

Nos anos 80 e 90 com avanços nas áreas tecnológicas e aparecimento de novas ideologias relativamente aquilo que seria a forma mais eficiente de relacionamento com o cliente, começam a aplicar-se modelos de *Machine Learning* e *frameworks de big data* para a modelação do CLV dos clientes [2].
De modo geral, quando se usam algoritmos de *Machine Learning* é necessário realizar o pré-processamento dos dados, onde se inclui a transformação, limpeza de dados e seleção de *features*. Os modelos de *machine learning*, neste contexto, podem ser utilizados com dois propósitos diferentes: 1) seleção das variáveis mais importantes; 2) construção de modelos preditivos do valor do CLV. A importância das variáveis utilizadas é tida em conta para identificar variáveis explicativas e identificar aquelas que maior influência têm sobre o *target* [35].
Os métodos para *feature selection* classificam-se em:

- ***Filter methods***: caracterizam-se por "filtrar" o *dataset* antes da construção do modelo de modo a ficar com um *subset* que contém as *features* mais relevantes para o resultado. Para escolher quais as variáveis usa-se, por exemplo, a correlação de Pearson.

- ***Wrapper methods***: alia a seleção de *features* com a construção de modelos. Caso haja um aumento do valor da *accuracy* ou outra medida de performance, a *feature* será adicionada, caso contrário será removida. Sabbeh (2018) aplicou esta metodologia no seu trabalho: usou o modelo *Random Forest* e fez a avaliação da relevância das *features* com *mean decrease accuracy*. Através da permuta de valores em cada uma das *features* permite medir o impacto de cada uma delas na precisão do modelo [49].

- ***Embedded method***: semelhante aos métodos *wrapper* mas, permitem analisar a contribuição de cada *feature* para o modelo, ou seja, tira proveito do seu próprio processo de seleção de variáveis e de seguida executa o "processo"de classificação de *features*. Esta metodologia foi usada nesta dissertação quando se fez a avaliação da importância das *features* de modo a decidir quais é que seriam ser tidas em conta na abordagem de *clustering*. A escolha deste método deveu-se ao facto de na literatura lhe serem apontados benefícios como: a alta precisão, a boa capacidade de generalização e a interpretabilidade. À semelhança daquilo que foi feito por Sabbeh (2018), também nesta dissertação se optou por utilizar *Random Forest* para fazer essa seleção.

Ao longo dos anos foram sendo desenvolvidas algumas variantes dos métodos acima descritos, por exemplo, o método mRMR, que se enquadra nos *filter methods*, faz a seleção de *features* recorrendo ao cálculo do erro de validação para diferentes conjuntos. Com esta avaliação consegue também escolher o número ideal de *features*. Outra das vantagens da utilização deste método é o facto de permitir fazer a comparação das *features* selecionadas para diferentes classificadores. Salienta-se que este método faz a seleção das *features* tendo em consideração algumas condições pré-definidas sobre o *target* (sendo esta a razão pela qual este método é classificado com *filter methods*) [38, 62]. Esta abordagem foi utilizada para selecionar as *features* comportamentais dos clientes que têm maior relevância para o CLV, para posteriormente as modelar por uma distribuição de *Poisson*.
Também no tratamento dos *missings values*, Shao *et al.* (2007) optaram por usar *filter methods*, decidindo excluir todas as variáveis que tivessem mais de 30% de *missing values*.
São vários os métodos de *Machine Learning* utilizados para o cálculo do CLV [52]. Khajvand & Tarokh (2011) utilizaram *clustering* para segmentar os clientes em grupos homogéneos e o método ARIMA (método de previsão baseado em séries temporais em estado não estacionário) para estimar o valor futuro do cliente de cada segmento com análise da tendência do valor do cliente, tendo em consideração a sazonalidade inerente ao processo [29]. Também nesta dissertação se optou por utilizar uma abordagem de *clustering*, já que esta permite fazer um agrupamento de clientes tendo em conta a sua similaridade e desse modo pode-se fazer uma análise descritiva de cada um dos grupos. O conhecimento mais aprofundado de cada grupo permitirá tomar decisões de marketing mais assertivas.
Alguns autores, Sabbeh (2018), Shirazi & Mohammadi (2019), Mauricio *et al.* (2016) decidiram recorrer a modelos de *Machine Learning* para efetuar previsões da propensão do cliente ao *churn*, pois este é um fator que em algumas indústrias apresenta um peso considerável no CLV do cliente [36, 49, 53]. Outros autores tentaram prever o momento da vida em que o cliente poderia ser mais propenso ao *churn*. De entre os modelos utilizados para este objetivo destaca-se o modelo CART e a implementação de um sistema linear clássico de análise de modelos. No entanto, na literatura também existem relatos da utilização das CART com outros objetivos, como, por exemplo, a modelação direta do CLV, as razões apontadas para esta escolha são a flexibilidade e o facto de serem interpretáveis uma vez

que podem facilmente ser convertidas para linguagem natural, o que neste estudo pode ser entendido como uma mais-valia para ter entender a "relação" entre as variáveis e o modo como as suas interligações interferem no CLV do cliente [49]. A razão apontada por Hasheminejad (2018) para a implementação das CART foi o facto de serem útil para a gestão de clientes bancários [24], como o *dataset* utilizado neste estudo também provêm de uma instituição bancária considerou-se que este modelo responderia de forma adequada as particularidades deste negócio. Com esta técnica conseguiu-se determinar os clientes que pertenciam a cada uma das classes em função das variáveis preditivas [24].
Outro aspeto de enorme relevância para as empresas é a identificação dos fatores que contribuem para o *churn*, bem como os fatores que contribuíram para o lucro da empresa. Mauricio *et al.* (2016) optou por usar a regressão linear múltipla para identificar os fatores associados ao cliente que contribuem para o lucro da empresa, bem como para prever a vida útil do cliente antes do *churn* [36]. Posteriormente usa uma rede neuronal para estimar a contribuição futura do cliente e Cadeias de Markov para estabelecer os estados de transição. Foram vários os autores que se preocuparem em prever como iria ser a evolução do CLV dos seus clientes e com esse intuito recorreram às cadeias de Markov, que são um modelo probabilístico e por isso, permitem ter uma ideia da probabilidade de transição entre estados (definidos anteriormente). Salienta-se ainda que os modelos de Markov podem ser úteis para processos de tomada de decisão [16]. Em setores que envolvam relacionamento com o cliente, como, por exemplo o setor bancário, esta abordagem para prever o CLV tem suscitado interesse, pelo facto de apresentar bastante flexibilidade [16]. Na revisão da literatura foram vários os fatores considerados para definir os estados: possibilidade de mudança de marca, frequência de transações, recência. Depois de estudar as várias alternativas tendo sempre em consideração o modelo de negócio subjacente decidiu-se estudar o CLV utilizando cadeias de Markov dadas as vantagens que este modelo oferece e que já foram apresentadas. À semelhança daquilo que foi feito por Haenlein (2007), nesta dissertação também foram usadas variáveis relacionadas com o tipo de produtos que o cliente adquiriu e a "intensidade de uso" associada ao produto para definir os estados[22]. Salienta-se que outra das vantagens da utilização das cadeias de Markov na modelação do CLV é que podem ser tiradas algumas inferências sobre o comportamento do cliente [21].
O algoritmo *AdaBoost* (Adaptative Boosting) foi utilizado em 2 vertentes: 1) melhorar a performance de outros algoritmos de *Machine Learning*; 2) previsão do *churn* do cliente. O *AdaBoost* apresenta como principais vantagens o facto de ser rápido, simples, fácil de programar podendo, por isso, ser combinado de forma flexível com qualquer outro método para encontrar hipóteses fracas. Outra das grandes vantagens deste algoritmo é o facto de não ser propenso ao *overfitting* e *noise sensitive* possuindo, por isso, uma melhor capacidade de generalização [52].
Para aumentar a sua rentabilidade as empresas têm-se esforçado em fazer vendas de produtos *cross-selling* aos seus clientes mais fieis e lucrativos. Com esse intuito Hosseni & Tarokh (2011) desenvolveu uma *framework* que otimiza o valor corrente dos clientes

e a probabilidade de *churn* usando a Regressão Logística [25]. Neste modelo foi feita a classificação de clientes como fieis/infiéis para de seguida serem segmentados com base no seu valor corrente e lealdade. Os segmentos mais lucrativos foram depois selecionados para os modelos de *cross selling*.
Recentemente, alguns autores optaram por recorrer a redes neuronais para estimar o CLV. A *rede neural de Kohonen* (que corresponde a um SOM otimizado), foi utilizada por Mauricio (2016) e Ayoubi (2016) na previsão do CLV. O modelo teve como *input* o "modelo" WRFM e como output o valor para cada um dos clusters. Uma das vantagens de trabalhar com esta metodologia é o facto de ela permitir uma redução das dimensões uma vez que durante o processo alguns *clusters* são agrupados obtendo-se assim clusters otimizados. A principal desvantagem é a possibilidade de aparecerem *clusters* divididos, sendo nessa situação, difícil determinar quais os pesos atribuídos a cada um dos *inputs* [3, 36].
Foram várias as técnicas de *machine learning* que ao longo dos últimos anos foram aplicadas na tentativa de obter uma melhor precisão na previsão do CLV, tendo isso em mente decidiu-se fazer um estudo comparativo entre a *performance* dos seguintes algoritmos: CART, *Support Vector Machines* (SVM), SVM utilizando *Sequential Minimal Optimization* (SMO), *Additive Regression*, método *K-Star* e MLP. Verificou-se que o modelo que obteve uma melhor *accuracy* foi MLP [14]. Como um dos objetivos desta dissertação é a previsão do CLV com a máxima precisão e como no estudo comparativo referido o modelo MLP foi aquele que obteve melhores resultados decidiu-se também optar pela implementação desse modelo nesta dissertação.

2.2.3 *Big Data*

Recentemente recorreu-se a *frameworks de big data* para tentar aumentar o valor vitalício do cliente para a empresa. Sun *et al.* (2014) descreve no seu trabalho, a *iCARE*, *framework de big data* que permite uma análise inteligente do comportamento do cliente num ambiente de negócio bancário com o objetivo de encontrar clientes que tinham uma grande probabilidade de se tornar inativos no futuro [56]. As principais vantagens desta *framework* é o facto de os modelos serem implementados de maneira paralela (*map-reduce*) e escalável (o sistema tem capacidade de expansão sem perdas do seu desempenho). Esta arquitetura permite alcançar alto desempenho com baixo tempo de resposta na avaliação de modelos. A principal novidade deste trabalho foi o facto de ter usado dados não estruturados. Para tratar este tipo de dados foi usado a plataforma*big insights*, desenvolvida em Hadoop.
Outra das necessidades dos tempos atuais é a de ter informações/dados serem fornecidos em tempo real e de maneira sistemática. Vieira & Sehgal (2017) teve essa questão em consideração ao criar um algoritmo de recomendação [59]. Neste tipo de algoritmos são inferidos os itens em que um cliente poderá estar interessado, usando como input as preferências passadas (por exemplo, dados transacionais, *ratings*, monitorizações do comportamento do cliente através quantidade de aplicações baixadas ou sites visitados).

De modo a armazenar dados estruturados e não estruturados usou-se *Hadoop* e o *Spark*. Depois "utilizou-se" *collaborative deep learning*. Dado que esta abordagem aprende as representações profundas dos conteúdos enquanto considera a matriz *rating*, verificam-se melhorias na capacidade de extrapolação de *features* e poder preditivo do modelo.

2.2.4 Outras Abordagens

Alguns autores sugeriram abordagens "alternativas" para abordar a previsão do CLV. Por exemplo, Lycett & Marshan (2017) optaram por fazer a previsão do CLV tendo em consideração a rede de relacionamento dos clientes [35]. As *Social Network Analysis* (SNA) investigam as estruturas sociais existentes num conjunto de entidades, bem como as relações existentes entre essas entidades. Esta análise permite efetuar mapeamentos, medições de relacionamentos e fluxo de valores entre as entidades que formam a rede. Numa SNA os nós correspondem a grupos e organizações, enquanto que as arestas representam os relacionamentos entre os nós. Lycett & Marshan (2017) investigaram se os clientes que estão mais conectados tem uma influência maior sobre o *cluster*, organizações com as quais negoceiam e qual o fluxo de caixa gerado pelos clientes de primeira ordem, com os quais contactam. Para classificar os clientes como mais ou menos relevantes, várias medidas de centralidade foram consideradas, nomeadamente: centralidade de grau, que diz respeito ao número de arestas anexadas a um nó. Pode ser medida através do número de arestas que chegam a um determinado nó (*in-degree*) ou através do número de arestas que são iniciadas em determinado nó (*out-degree*). Um determinado nó é considerado central se por ele passar o caminho mais curto para todos os outros nós numa rede. A identificação de nós centrais pode melhorar as comunicações com uma rede. A centralidade de intermediação quantifica a importância do nó, enquanto que a centralidade do vetor próximo, mede a importância do nó através centralidade de nós conectados a ele.
Os programas de referência (fornecem incentivos financeiros aos clientes que tragam novos clientes para a empresa) também podem ser vistos como um bom indicador da lealdade do cliente, sendo por isso algo a ter em conta no cálculo do CLV, pressupõe-se que determinado cliente só recomende determinada companhia se estiver satisfeito com o serviço que esta lhe presta. Schmitt *et al.* (2011) fizeram um estudo sobre a relação entre o valor do cliente e os programas de referência recorrendo ao modelo de Cox [51]. Os programas de referência tem como principal desvantagem o facto de alguns clientes mudarem para a empresa em questão apenas pelos incentivos financeiros, desvinculando-se da empresa após terem recebido o valor monetário inerente ao programa. Esta abordagem assenta no pressuposto de hemofilia (tendência para pessoas se relacionarem com pessoas parecidas consigo). As principais conclusões deste estudo é que esta é uma maneira popular de adquirir clientes, não havendo, no entanto evidências que os clientes adquiridos por estes programas sejam mais valiosos que os restantes. Os clientes que chegam à empresa através de programas de referência tem uma margem de contribuição alta (embora essa diferença seja desgastada com o tempo). A métrica usada neste tipo programas é o valor

dos clientes adquiridos.

Aeron *et al.* (2012) sugeriram a utilização de algoritmos genéticos para automatizar a tomada de decisão em dois processos: seleção de variáveis e na escolha de segmentos ideais com base no CLV. Nesta *framework* é usado *clustering* para segmentar e algoritmos genéticos para otimizar a segmentação [1].

Dada a competitividade existente nas diferentes indústrias cria-se a necessidade de em todas elas se apostar na personalização dos produtos oferecidos aos clientes. Galal *et al.* (2016) apresentaram uma *framework* para controlar o processo de tomada de decisão, aprimorar a automatização do processo de personalização e melhorar a eficácia da segmentação [20]. O objetivo deste projeto era desenvolver cartões e programas de recompensa personalizados diferentes características do cliente. De modo a analisar a eficácia desta abordagem os clientes foram divididos em dois conjuntos: no primeiro conjunto todos os clientes foram abrangidos pela mesma campanha de *marketing* sem ter feita nenhuma personalização (GBO) enquanto que na segunda abordagem foram construídos perfis personalizados pelas unidades de negócio com base nas pesquisas de mercado. Os perfis foram traçados com recurso a árvores de decisão, dado o seu formato intuitivo permite descrever o perfil de cada segmento.

A segmentação dos clientes pode ser feita com base em diferentes critérios, por exemplo, na segmentação por benefícios a identificação dos segmentos de mercado é feita com base por fatores causais. Por exemplo, usando o motivo da preferência como *feature*, podem ser descobertos benefícios diferentes para diferentes grupos de clientes [28].

Capítulo 3

Dataset

Neste capítulo é feita a descrição e caracterização do dataset usado nesta dissertação.

3.1 Descrição e caracterização do *dataset*

A seleção de variáveis para os modelos de predição foi feita com base em referências existentes no estado da arte, apesar de terem sido feitas algumas adaptações, uma vez que os dados usados nesta dissertação provêm de uma instituição bancária e alguns dos artigos analisados utilizam dados dados provenientes de outros ramos de negócio. No *dataset* utilizado existem quer variáveis monetárias, que são aquelas que têm um valor associados (em euros) e por isso, são quantitativas, quer variáveis não monetárias, que estão associadas ao estilo de vida e padrão de comportamento do cliente. Podem ser variáveis quantitativas ou qualitativas. De forma resumida, as variáveis existentes no *dataset* podem ser classificadas em 5 classes diferentes:

- Demográficas: Variáveis que caracterizam a amostra quanto ao grau de escolaridade, idade, situação profissional, ect;
- Transações: Podem-se subdividir em depósitos e em transferências;
- Fidelização: Estas variáveis reflectem o grau de afinidade existente entre o cliente e a empresa;
- Posse: Permitem descrever os produtos bancários que o cliente possui;
- Reclamações: Variáveis que servem para fazer uma análise das reclamações feitas pelo cliente;

Além das variáveis descritas na tabela 3.1, foram também incluídas, em alguns casos, as suas variáveis homólogas, relativas ao ano anterior, ou aos dois anos anteriores (no caso do valor da rentabilidade líquida). A inclusão destas variáveis foi uma sugestão de especialistas do negócio bancário, por acreditarem que iriam dar bons *insights* para a previsão do CLV. As variáveis em que se considerou o valor no ano anterior foram: número total de depósitos, valor total de depósitos, número de depósitos em ATM, valor dos depósitos em ATM, número de depósitos no balcão, valor dos depósitos no balcão, valor total de transferências a crédito, número total de transferências a crédito, valor total de transferências a crédito, valor total de compras a crédito, número total de compras a crédito, número de visitas ao site, posse de produtos *cross selling*, rentabilidade líquida, número de simulações e valor de simulações. Na tabela 3.1 resume-se as variáveis incluídas no *dataset*.

3.2 Pré-processamento dos dados

Antes de começar a análise preditiva, o *dataset* deve ser preparado de forma apropriada. O objetivo do pré-processamento de dados é assegurar a qualidade da análise.
Nesta etapa procedeu-se ao tratamento dos *missing values*. Nos casos em que os *missings* significavam que o cliente não possuía o produto, foram substituídos por 0. Nos outros casos colocou-se NaN (*"not a number"*) de forma a identificar casos omissos.
Para além disso, nesta etapa também foram construídas novas variáveis, recorrendo ao domínio do conhecimento da área em estudo, com o objetivo de aumentar a capacidade preditiva dos modelos construídos posteriormente. Este processo denomina-se de *feature engineering*. Na tabela 3.2 apresentam-se as *features* criadas bem como as fórmulas usadas para o seu desenvolvimento.

Também se procedeu à identificação e exclusão de *outliers*, dado que estes são valores atípicos podendo causar anomalias nos resultados obtidos por meio de algoritmos e sistemas de análise. Assim foram removidos os valores que fossem 1.5 vezes superiores á diferença entre o terceiro quartil, Q3, e o primeiro quartil Q1, obtendo-se para os *outliers* inferiores e superiores, respetivamente,

$$\begin{aligned} &Q1 - 1.5 \times (Q3 - Q1) \\ &Q3 + 1.5 \times (Q3 - Q1). \end{aligned} \tag{3.1}$$

Tabela 3.1: Descrição das variáveis

Designação	Classe	Descrição	Tipo	Modelos em que foi utilizada	Referências no estado da arte
Idade do cliente	Demográficas		Discreta	RF, CART, MLP, DTMC	[39], [28], [37], [3]
Número de dependentes	Demográficas	Número de pessoas que dependem diretamente do cliente	Discreta	RF, CART, MLP	[43], [52]
Número de relações familiares	Demográficas	Quantifica o número de familiares do cliente que também tem uma conta ativa no banco	Discreta	RF, CART, MLP	[39], [28], [37]
Sexo	Demográficas		Binário	RF, CART, MLP, DTMC	[39], [28], [37], [22], [36], [3]
Estado civil	Demográficas		Nominal	RF, CART, MLP, DTMC	[39], [28], [22], [36], [3]
Grau de instrução	Demográficas		Ordinal	RF, CART, MLP, DTMC	[39], [28], [37], [3] [52]
Profissão	Demográficas		Nominal	RF, CART, MLP	[39], [28], [37]
Situação profissional	Demográficas		Nominal	RF, CART, MLP	[39], [28], [37]
Número total de depósitos	Tansações	Número total de depósitos efetuados por cada cliente por mês	Discreta	RF, CART, MLP	[39], [1], [37], [36], [56]
Número de depósitos no balcão	Tansações	Número total de depósitos efetuados por cada cliente por mês num balcão	Discreta	RF, CART, MLP	[39], [1], [56]
Número de depósitos em ATM	Tansações	Número total de depósitos efetuados por cada cliente por mês num ATM	Discreta	RF, CART, MLP, DTMC	[56], [22]
Número total de transferências a crédito	Tansações	Número total de transferências efetuados por cada cliente por mês	Discreta	RF, CART, MLP	[39], [1], [36], [56]
Número total de compras a crédito	Tansações	Número total de transferências efetuados por cada cliente por mês em compras	Discreta	RF, CART , MLP	[39], [43], [1], [36], [56]

Tabela 3.1: Descrição das variáveis

Designação	Classe	Descrição	Tipo	Modelos em que foi utilizada	Referências no estado da arte
Número de transações por iniciativa própria	Transações	Número de transações por iniciativa do cliente por mês	Discreta	RF, CART, DTMC, MLP	[22]
Número de levantamentos	Transações	Número de levantamentos efetuados pelo cliente por mês	Discreta	RF, CART, DTMC, MLP	[22]
Número de outros depósitos	Transações	Número de "outros" depositos efetuados pelo cliente	Discreto	RF, CART, MLP	[22]
Valor dos depósitos no balcão	Tansações	Montante total(em euros) dos depósitos efetuados por cada cliente por mês num balcão	Contínua	RF, CART, MLP	[39], [56]
Valor dos depósitos em ATM	Tansações	Montante total(em euros) dos depósitos efetuados por cada cliente por mês num ATM	Contínua	RF, CART, MLP	[39], [1], [56]
Valor total das transferências a crédito	Tansações	Montante total (em euros) das transferência efetuados por cada cliente por mês	Contínua	RF, CART, MLP	[39], [3], [56]
Valor total das compras a crédito	Tansações	Montante total (em euros) das transferência efetuados por cada cliente por mês em compras	Contínua	RF, CART, MLP	[39], [43], [56]
Valor total de depósitos	Tansações	Montante total (em euros) dos depósitos efetuados por cada cliente por mês	Contínua	RF, CART, MLP	[39] [3], [56]
Número de simulações	Fidelização	Número total de simulações de compra de produtos feitas por cada cliente em cada mês	Discreta	RF, CART, MLP	[20]
Visitas ao site	Fidelização	Número total de visitas ao site feitas por cada cliente, por mês	Discreto	RF, CART, MLP	[1]
Antiguidade	Fidelização	Antiguidade da relação do cliente com a instituição bancária	Discreto	RF, CART, MLP	[59]

Tabela 3.1: Descrição das variáveis

Designação	**Classe**	**Descrição**	**Tipo**	**Modelos em que foi utilizada**	**Referências no estado da arte**
Número de contas ativas	Fidelização	Número de contas que o cliente tem e "movimenta" na instituição bancária	Discreto	RF, CART, MLP	[59]
Número de idas à sucursal	Fidelização	Número total de idas à sucursal por cliente em cada mês	Discreta	RF, CART, MLP	[1]
NPS *score*	Fidelização	*Net Promotor Score*. Métrica usada para medir a lealdade do cliente	Contínua	RF, CART, MLP	[37]
Valor das simulações	Fidelização	Quantias (em euros) envolvidas nas simulações de aquisição de produtos	Contínua	RF, CART, MLP	[20]
Ativo	Fidelização	*Flag* que assume o valor 1 quando o cliente possui produtos que paga com uma certa periodicidade ou tem saldo de conta corrente superior a 100 euros. Caso esta condição não se verifique o valor da *flag* será 0.	Binária	RF, CART, DTMC, MLP	[22]
Número de cartões de crédito *Visa* & *Mastercard*	Posse	Quantidade de cartões de crédito *Mastercard* que o cliente possui	Discreta	CART, RF, DTMC, MLP	[22]
Número de cartões de crédito AM EXP	Posse	Quantidade de cartões de crédito *American Express* que o cliente possui	Discreta	CART, RF, DTMC, MLP	[22]
Rentabilidade líquida dos últimos 12 meses	Posse	Rentabilidade Líquida do cliente no último ano	Contínua	RF, CART, MLP, DTMC	[43], [39], [28], [37], [56], [33]
Montante em produtos *cross-selling*	Posse	Montante (em euros) resultante da aquisição de produtos *cross-selling*	Contínua	RF, CART, MLP	[25], [56]

Tabela 3.1: Descrição das variáveis

Designação	Classe	Descrição	Tipo	Modelos em que foi utilizada	Referências no estado da arte
Valor dos recursos a prazo	Posse	Valor (em euros) dos recursos a prazo que o cliente possui	Contínua	RF, CART, MLP, DTMC	[28], [56], [22]
Valor dos recursos à ordem	Posse	Valor (em euros) dos recursos à ordem que o cliente possui	Contínua	RF, CART, MLP	[28], [56]
Valor dos recursos a título	Posse	Valor (em euros) dos recursos a título que o cliente possui	Contínua	RF, CART, MLP	[28]
Valor de outros recursos	Posse	Valor (em euros) de outros recursos que o cliente possui	Contínua	RF, CART, MLP	[28], [56]
Valor do seguro de vida	Posse	Valor (em euros) pago pelo seguro de vida	Contínua	RF, CART, MLP, DTMC	[22]
Valor do crédito vivo	Posse	Valor do crédito que o cliente ainda têm em dívida com a instituição bancária	Contínua	RF, CART, MLP	[28], [59]
Saldo da conta corrente	Posse	Valor médio do saldo existente na conta á ordem durante o intervalo de tempo considerado neste estudo	Contínua	RF, CART, MLP, DTMC	[22]
Valor do crédito habitação	Posse	Montante (em euros) do crédito à habitação em dívida	Contínua	CART, RF, MLP DTMC	[22]
Valor do crédito pessoal	Posse	Montante (em euros) do crédito pessoal em dívida	Contínua	CART, RF, MLP, DTMC	[22]
Valor do seguro automóvel	Posse	Valor (em euros) pago pelo seguro automóvel	Contínua	RF, CART, MLP, DTMC	[22]
Valor do seguro de saúde	Posse	Valor (em euros) pago pelo seguro de saúde	Contínua	CART, RF, MLP, DTMC	[22]
Valor dos seguros de risco não vida	Posse	Valor (em euros) pago pelo seguros de risco. Nesta variável não está a ser considerado o seguro de vida	Contínua	CART, RF, MLP, DTMC	[22]

Tabela 3.1: Descrição das variáveis

Designação	Classe	Descrição	Tipo	Modelos em que foi utilizada	Referências no estado da arte
Valor do seguro de acidentes pessoais	Posse	Valor (em euros) pago pelo seguro de acidentes pessoais	Contínua	CART, RF, MLP, DTMC	[22]
Valor do seguro de multirriscos	Posse	Valor (em euros) pago pelo seguro de multirriscos	Contínua	CART, RF, MLP, DTMC	[22]
Valor de outros seguros	Posse	Valor (em euros) pago por outros seguros que o cliente possui	Contínua	CART, RF, MLP, DTMC	[22]
Valor da poupança à habitação	Posse	Valor (em euros) existente na conta poupança para habitação e condomínio	Contínua	CART, RF, MLP, DTMC	[22]
Valor de outras poupanças	Posse	Valor (em euros) existente em outras contas poupança	Contínua	CART, RF, MLP, DTMC	[22]
Valor da poupança reforma	Posse	Valor (em euros) existente na conta poupança reforma	Contínua	RF, CART, MLP, DTMC	[22]
Valor do património financeiro	Posse	Património financeiro do cliente	Contínua	RF, CART, MLP	[43], [39], [37], [22]
Número de dias de resolução	Reclamações	Número de dias necessários para a reclamação ser resolvida	Discreto	RF, CART, MLP	[33]
Valor reclamado	Reclamações	Quantia (em euros) exigida pelo cliente quando faz a reclamação	Contínua	RF, CART, MLP	[33]
Valor estornado	Reclamações	Quantia (em euros) reembolsada ao cliente após uma reclamação, quando lhe é dada razão	Contínua	RF, CART, MLP	[33]
Razão ao cliente	Reclamações	*Flag* que indica se foi ou não reconhecida razão ao cliente pela sua reclamação	Binária	RF, CART, MLP	[33]

CART: *Classification and Regression Trees*, **RF**: *Random Forest*, **MLP**: *Multi Layer Perceptron*, **DTMC**: *Discrete time Markov chains*

Tabela 3.2: Variáveis criadas

Variável	**Definição da fórmula de cálculo**
Taxa de variação do número total de depósitos.	Razão cujo numerador corresponde à diferença obtida pelo número total de depósitos feitos este ano e o número total de depósitos feitos no ano anterior e o denominador corresponde ao número total de depósitos feitos este ano.
Taxa de variação do valor total dos depósitos.	Razão cujo numerador corresponde à diferença obtida pelo valor total dos depósitos feitos este ano e o valor total de depósitos feitos no ano anterior e o denominador corresponde ao valor total de depósitos feitos este ano.
Taxa de variação do número de depósitos em ATM.	Razão cujo numerador corresponde à diferença obtida pelo número total de depósitos feitos este ano em ATM e o número total de depósitos feitos no ano anterior em ATM e o denominador corresponde ao número total de depósitos feitos este ano em ATM.
Taxa de variação do valor dos depósitos em ATM	Razão cujo numerador corresponde à diferença obtida pelo valor total dos depósitos feitos este ano em ATM e o valor total de depósitos feitos no ano anterior em ATM e o denominador corresponde ao valor total de depósitos feitos este ano em ATM.
Taxa de variação do número de depósitos no balcão.	Razão cujo numerador corresponde à diferença obtida pelo número total de depósitos feitos este ano no balcão e o número total de depósitos feitos no ano anterior no balcão e o denominador corresponde ao número total de depósitos feitos este ano no balcão.
Taxa de variação do valor dos depósitos no balcão.	Razão cujo numerador corresponde à diferença obtida pelo valor total dos depósitos feitos este ano no balcão e o valor total de depósitos feitos no ano anterior no balcão e o denominador corresponde ao valor total de depósitos feitos este ano no balcão.

Tabela 3.2: Variáveis criadas

Variável	**Definição da fórmula de cálculo**
Taxa de variação do valor total de transferências a crédito.	Razão cujo numerador corresponde à diferença obtida pelo valor total de transferências a crédito feitas este ano e o valor total de transferências a crédito feitas no ano anterior e o denominador corresponde ao valor total de transferências feitas este ano.
Taxa de variação do número total de transferências a crédito.	Razão cujo numerador corresponde à diferença obtida pelo número total de transferências a crédito feitas este ano e o número total de transferências a crédito feitas no ano anterior e o denominador corresponde ao número total de transferências feitas este ano.
Taxa de variação do valor de compras a crédito.	Razão cujo numerador corresponde à diferença obtida pelo valor total de compras a crédito feitas este ano e o valor total de compras a crédito feitas no ano anterior e o denominador corresponde ao valor total de compras a crédito feitas este ano.
Taxa de variação do número de compras a crédito.	Razão cujo numerador corresponde à diferença obtida pelo número total de compras a crédito feitas este ano e o número total de compras a crédito feitas no ano anterior e o denominador corresponde ao número total de compras a crédito feitas este ano.
Taxa de variação do número de *logins* no site.	Razão cujo numerador corresponde à diferença obtida pelo número total de *logins* feitos este ano e o número total de *logins* feitos no ano anterior e o denominador corresponde ao número total de *logins* feitos este ano.
Taxa de variação do valor de simulações.	Razão cujo numerador corresponde à diferença obtida pelo valor total das simulações feitos este ano e o valor total das simulações feitas no ano anterior e o denominador corresponde ao valor total das simulações feitas este ano.
Taxa de variação do número de simulações.	Razão cujo numerador corresponde à diferença obtida pelo número total das simulações feitos este ano e o número total das simulações feitas no ano anterior e o denominador corresponde ao número total das simulações feitas este ano.

Tabela 3.2: Variáveis criadas

Variável	**Definição da fórmula de cálculo**
Taxa de variação do montante de produtos (*cross-selling*).	Razão cujo numerador corresponde à diferença obtida pelo valor dos produtos *cross selling* que o cliente possui este ano e o valor dos produtos *cross selling* que o cliente possuía no ano passado e o denominador corresponde ao valor dos produtos *cross selling* que o cliente possui este ano.
Taxa de variação da rentabilidade líquida dos últimos 12 meses.	Razão cujo numerador corresponde à diferença obtida pelo valor da rentabilidade líquida que o cliente possui este ano e o valor da rentabilidade líquida que o cliente possuía no ano passado e o denominador corresponde ao valor da rentabilidade líquida que o cliente possui este ano.

Capítulo 4

Métodos

Neste capítulo é feita uma apresentação teórica das técnicas utilizadas para modelar o Customer Lifetime Value.
Os dados utilizados nesta dissertação estavam armazenados em tabelas SAS, por esse motivo, a construção dos datasets e a etapa de feature engineering foram feitas com recurso ao software SAS [50]. A análise estatísticas dos dados, a construção de modelos preditivos e a análise de clusters foram efetuadas com o software R [41].

4.1 Análise exploratória e descritiva do *dataset*

A análise exploratória dos dados consiste na aplicação métodos, estatísticas e técnicas de visualização. O objetivo é descobrir padrões, características, tendências, comportamento anómalos e *outliers*, visando maximizar a obtenção de informações ocultas.
Para caracterizar cada uma das variáveis que constam no *dataset* foram utilizadas medidas de tendência de centralidade, medidas de dispersão e medidas de localização. As medidas de tendência de centralidade utilizadas foram: a **média ($\bar{x}$)**, que corresponde ao somatório de todos os elementos a dividir pelo número total de elementos; a **mediana ($\tilde{x}$)**, que corresponde ao valor da variável não inferior nem superior a metade dos valores observados, quando os valores foram dispostos por ordem crescente; o **mínimo (Min)**, que corresponde ao valor mínimo que a variável em análise assume e o **máximo (Máx)**, que corresponde ao valor máximo que a variável em análise assume. As medidas de dispersão utilizadas foram: o **desvio de padrão (DP)**, que fornece a medida de dispersão em relação á média. É obtido pelo cálculo da raiz quadrada da variância, e o **coeficiente de variação (CV)**, que é uma medida padronizada de dispersão de uma distribuição de probabilidade ou de frequências. Para calcula-lo, divide-se o desvio-padrão pela média. Os quartis, que são medidas de localização, são obtidos através da divisão da distribuição

em quatro partes iguais, sendo que, as observações foram previamente ordenadas por ordem crescente. Recorrendo ao **primeiro quartil (Q1)**, sabe-se que 25% das observações são menores que ou iguais a este valor. O **terceiro quartil (Q3)**, indica que 75% dos dados são menores que ou iguais a este valor.
A lógica subjacente à construção das representações gráficas foi a seguinte: Para as variáveis categóricas / discretas, que apresentavam uma pequena amplitude de valores foram construídas tabelas, que se encontram ordenadas por ordem decrescente da frequência associada, que resumem a percentagem de clientes em cada "categoria" / para cada valor; Para as variáveis discretas que apresentam uma grande amplitude de valores, recorreu-se a gráficos de barras para fazer a representação, sendo que, no eixo das ordenadas é representada a percentagem de clientes que existe em cada categoria; No caso das variáveis contínuas, optou-se por fazer a representação da respetiva distribuição usando *boxplots* e histogramas. Salienta-se que, apesar de nos histogramas se encontrarem representadas as curvas de densidade probabilística, posteriormente não é feita qualquer interpretação das mesmas, uma vez que em nenhuma das variáveis a curva de densidade ilustra a uma distribuição estatística conhecida. Nos *boxplots* que são apresentados na secção 5, os *outliers* foram retirados previamente com recurso à opção OUTLINE=FALSE do *package boxplot* do R [41].

4.2 CART

O modelo CART (*Classification and Regression Tree*) não é mais que uma árvore de decisão binária resultante de um conjunto de regras *if-then* que tanto podem ser usadas para problemas de classificação como para problemas de regressão. Este tipo de modelos tem a vantagem de permitir lidar com grandes quantidades de dados, apresentando poucas restrições e pressupostos. Além disso, permitem criar modelos precisos mesmo em bases de dados com ruído, dados inconsistentes ou incompletos, sendo um dos principais aspetos positivos o facto de ser de fácil interpretação e compreensão.
Neste tipo de modelos faz-se uma partição recursiva dos dados conseguindo-se assim descobrir padrões que, na maior parte dos casos, não são facilmente detectáveis. O princípio básico de indução de uma Árvore de Decisão é o famoso lema *"Dividir para Conquistar"*.
As árvores de decisão organizam-se à luz da estrutura de uma árvore, com nós, onde ocorrem as decisões, ramos, representativos das opções dos nós, e folhas, correspondentes aos resultados das decisões. Têm como ponto de partida o nó da raiz onde se encontra toda a amostra. Cada um dos nós contêm uma pergunta binária (com resposta sim/não) sobre alguma variável. À medida que a árvore vai crescendo, a divisão do conjunto de dados continua nos ramos inferiores da árvore, utilizando-se novas variáveis de partição e novos valores de corte. As folhas da árvore contêm uma classificação na decisão. As listas resultantes da decisão são uma forma reduzida da árvore, onde se consegue compreender quais as perguntas que levaram ao agrupamento das amostras nas folhas da árvore. As

folhas terminais da árvore correspondem ao valor médio do *target* (no caso deste ser contínuo). O processo de treino do modelo ocorre até atingir um critério de homogeneidade ou até que sejam satisfeitos os critérios de paragem, tratando-se, por isso, de um algoritmo de partição binário recursivo. Por este ser um método recursivo podem ocorrer problemas de instabilidade, ou perda do poder de generalização. Para evitar estes problemas foram usadas duas soluções: 1) **método da poda**: tem como objetivo a minimização das taxas de erro, em cada conjunto de validação. À medida que as árvores crescem, o poder de generalização dos nós terminais diminui, então espera-se que a eliminação de alguns destes nós terminais diminuía o erro, diminuindo o sobre ajuste e aumentando o poder de generalização da árvore; 2) utilização de ***Random Forest***: neste caso o *output* final resulta da combinação dos resultados de um conjunto de árvores. Verificou-se que a combinação de vários modelos fracos *(weak predictors)* gera um modelo combinado bastante robusto e preciso.

Para medir a homogeneidade de cada conjunto de dados usa-se o **Índice de Impureza**. A diminuição do índice de impureza é chamado de **Ganho de Informação** (GI).

Para a partição s, a variável X que obtiver o maior ganho de informação será a escolhida como variável de partição. Por outro lado, para cada possível variável de partição, X, o valor de corte no n-ésimo nó, X_n, será aquele que maximize o ganho de informação. O índice de impureza define-se pela variância do *target*, $Var(Y)$,

$$\begin{aligned} Var(Y) &= \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2 \\ \text{sendo } \mu &= \frac{1}{N}\sum_{i=1}^{N}y_i \end{aligned} \tag{4.1}$$

pois o valor da variância diminui à medida que os valores de Y ficam mais homogéneos, apresentando-se mais próximos da média.

O Ganho de Informação no n-ésimo nó da partição s é dado pela diferença entre a impuridade do nó pai e a impuridade dos nós filhos, isto é, representando p o nó pai e f os nós filhos[1], tem-se

$$GI(Y,s) = Var(Y)_p - Var(Y)_f \tag{4.2}$$

sendo $Var(Y)_p$ a impuridade associada ao *dataset* inicial, D, de dimensão N[2]. A impuridade dos nós filhos, $Var(Y)_f$, é dada pela média ponderada das variâncias de cada um dos nós filhos. Assumindo uma partição do *dataset* D, em dois subconjuntos D_L (L - *Left*)

[1] Salienta-se que f, que no capítulo 2 representava o número de compras efetuadas neste contexto passará a representar os nós filhos.

[2] Salienta-se que N, que no capítulo 2 representava o valor de *network* do cliente, neste contexto passará a representar a dimensão do *dataset* inicial.

e D_R (*R - Right*), com variâncias $Var(Y)_L$ e $Var(Y)_R$ e dimensões respectivamente iguais a N_L e N_R, a impuridade dos nós filhos é definida por

$$Var(Y)_f = \frac{N_L}{N} \times Var(Y)_L + \frac{N_R}{N} \times Var(Y)_R \tag{4.3}$$

Assim, o Ganho de Informação pode expressar-se pela equação

$$GI(Y,s) = Var(Y)_p - \left(\frac{N_L}{N} \times Var(Y)_L + \frac{N_R}{N} \times Var(Y)_R \right). \tag{4.4}$$

Quanto maior o Ganho de Informação, melhor será o critério de partição escolhido. O melhor ponto de corte da variável X, no n-ésimo nó da partição s será aquele que maximize o ganho de informação em s, ou seja

$$X_n = \underset{s}{argmax}\Big(GI(Y,s) \Big) \tag{4.5}$$

Neste estudo, a CART foi implementada utilizando o *package rpart* [58] do R [41]). Para calcular a importância de cada uma das variáveis recorreu-se à função *gain.ratio* do *package FSelector* [47] do R [41]. Esta função permite encontrar o pesos das variáveis de *input* com base em sua correlação que tem com o *target*. A métrica utilizada na função *gain.ratio* para fazer esse cálculo foi o a entropia de Shannon, que de forma genérica pode ser entendida como a quantidade de informação armazenada em cada variável. Assim, o cálculo da importância de cada variável tem em conta três componentes: a entropia de Shannon para o *target*, $H(Target)$; a entropia condicional de Shannon para cada uma das variáveis de *input* do modelo e a variável *target*, $H(Target, Attribute)$, e entropia de Shannon para cada uma das variáveis de *input* do modelo,$H(Attribute)$. Matematicamente o *gain.ratio* é traduzido por [40]:

$$\frac{H(Target) + H(Attribute) - H(Target, Attribute)}{H(Attribute)}. \tag{4.6}$$

Com o objetivo de encontrar o tamanho ideal da árvore procedeu-se à poda (*prunning*) das árvores geradas. A poda é uma técnica utilizada para reduzir o tamanho das árvores de decisão, removendo os nós menos relevantes para a predição do *target*. A poda reduz a complexidade, melhora a precisão preditiva e, reduz a probabilidade de ocorrência de *overfitting*. No *package rpart* [58], o controlo do tamanho da árvore é feito pelo parâmetro de complexidade (cp), que impõe penalidades caso a árvore sofra múltiplas divisões. O tamanho da árvore é inversamente proporcional ao valor do *complexity parameter* (cp), ou seja, um cp demasiado baixo, irá produzir árvores grandes e, que se encontram mais propensas ao *overfitting*, e um cp demasiado elevado irá produzir árvores pequenas e, por isso, demasiado generalistas. Em ambos os casos há um decréscimo da performance preditiva do modelo. O *maxdeph* é outro parâmetro do *package rpart* que pode ser usado para evitar que sejam criadas árvores com demasiada profundidade, uma vez que este

parâmetro permite para definir a profundidade máxima de qualquer nó da árvore final. O *minsplit* foi outro dos parâmetros usados na construção das CARTs. Com este parâmetro é possível definir o número mínimo de observações que devem existir num nó antes de se tentar fazer a divisão naquele nó especifico, evitando assim, que sejam feitas divisões tendo por base um número reduzido de amostras.

4.3 Random Forest

As árvores de decisão apesar de serem um modelo de fácil implementação e compreensão, apresentam como desvantagem o facto de serem propensas ao *overfitting*. As Random Forest (RF) oferecem uma solução a esse problema através da combinação de múltiplas árvores aleatórias, uma vez que a substituição de uma árvore por um conjunto delas aumenta o poder de generalização [9]. Para cada árvore, o algoritmo selecciona de forma aleatória amostras de dados com o objetivo de construir árvores independentes. A aleatoriedade necessária para a construção de árvores independentes a partir dos mesmos dados de treino pode ser conseguida utilizando, por exemplo, uma abordagem de ***bagging***, tal como Breiman (2001) propõe [9]. O *bagging* consiste na combinação das técnicas *bootstrap* com *agregação*, ou seja, cada elemento da amostra é selecionado de forma aleatória com ou sem reposição de acordo com uma distribuição uniforme. Por fim as previsões geradas aleatoriamente por cada árvore são agregadas usando a média.
A *Random Forest* define-se por um conjunto de árvores de decisão independentes.
A flexibilidade das árvores de decisão deve-se ao facto de estas possuírem dois parâmetros, que influenciam os graus de liberdade: o número de árvores e a "profundidade" das árvores (*tree deep*). O aumento do número de árvores permite que haja uma diminuição média do ruído das árvores, e consequentemente o erro de predição diminua. A profundidade da árvore afeta diretamente a capacidade de generalização da árvore: uma árvore pequena não apresenta muita confiança nas suas previsões devido ao facto de ainda existirem muitos dados heterogéneos agrupados nas suas folhas, no entanto, uma árvore demasiado grande terá poucos dados agrupados nas suas folhas, dificultando o cálculo estatístico fiável. Em síntese, a árvore até determinado momento está a aprender a fazer um bom ajuste das variáveis, mas a partir do momento em que a árvore começa a explicar demasiado bem o *training set*, perde poder de generalização.
Considerando uma árvore F que constituí a *Random Forest*, começa-se por se efetuar uma partição, s_t, no *dataset* inicial, D, tal como acontece nas CART. Cada árvore associa a uma observação w (tal que $w \in X$) a um sub conjunto de dados C_t, resultante da partição s_t. A RF "inteira" irá corresponder a uma "função" que associa w a um conjunto de dados:

$$A(w) = \left\{C_1, ..., C_f, ..., C_F\right\}. \tag{4.7}$$

Caso se considere que cada partição é equiprovável, a predição da RF pode ser calculada através da média das árvores posteriores,

$$P(Y|w) = \frac{1}{F}\sum_{t=1}^{F} P(Y|w \in C_f, s_t). \tag{4.8}$$

As RF podem ser regressivas ou de classificação dependendo do *output* desejado: no caso das árvores de regressão o *target* é contínuo, enquanto que nas RF de classificação é categórico. Ambos permitem modelar de forma eficiente funções não lineares, são escaláveis para um grande conjunto de dados de treino e grandes espaços multidimensionais de

input e *output*. Nesta dissertação foram utilizadas RF regressivas para a predição do CLV, geradas usando a implementação do *package randomForest* [34], do R [41].
Na otimização das *Random Forest* foram tidos em conta 3 parâmetros: o número de árvores (*ntree*), que diz respeito ao número de árvores utilizadas na previsão; o tamanho mínimo dos nós terminais (*nodesize*), que estabelece o número mínimo de amostras que tem obrigatoriamente de existir em cada folha e o *maxnodes*, que estabelece o número máximo de folhas por árvore.
Caso o *nodesize* apresente um valor elevado, isso irá condicionar o tamanho das árvores, que acabarão por ficar mais pequenas e, consequentemente, o tempo de processamento irá ser menor.
Caso não se defina nenhum valor para o número máximo de nós terminais que as árvores podem ter (*maxnodes*), as árvores irão crescer o máximo que lhes for possível, sendo apenas limitadas pelo parâmetro *nodesize*, caso tenha sido definido.
As variáveis mais importantes na predição do *output* foram determinadas recorrendo à função *importance* do *package randomForest* [34]. Para determinar a importância de cada variável, esta função utiliza a %IncMSE, que é calculada a partir da permutação de dados OOB (*out-of-bag*) para cada variável preditora. Neste caso e, como estamos perante uma regressão, o algoritmo calcula o *Mean Squared Error*.

4.4 Cadeias de Markov

As cadeias de Markov correspondem a um processo estocástico com estados discretos (*Discrete Time Markov Chain*, DTMC) que apresentam inúmeras aplicações práticas. Permitem modelar probabilidades de transição entre estados discretos recorrendo a matrizes.
A DTMC pode representar-se por uma sequência de variáveis aleatórias X_1,X_2,...,X_n,... caracterizada pela propriedade de Markov, que estabelece que, a distribuição do estado seguinte (X_{n+1}) depende apenas do estado atual (X_n), não apresentando qualquer dependência dos estados anteriores ($X_{n-1}, X_{n-2}, .., X_1$), isto é,

$$P(X_{n+1} = x_{n+1} | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x_{n+1} | X_n = x_n). \tag{4.9}$$

O conjunto de estados possíveis de X_n, finito ou contável, denomina-se de espaço de estados da cadeia. A probabilidade p_{ij} [3] de se mover / transitar de um estado i para outro j é denominada por probabilidade de transição:

$$p_{ij} = P(X_1 = s_j | X_0 = s_i). \tag{4.10}$$

Assim, a probabilidade de transição do estado i para o estado j no n-ésimo passo, $p_{ij}^{(n)}$, é denotada por

$$p_{ij}^{(n)} = P(X_n = j | X_{n-1} = i). \tag{4.11}$$

[3] Salienta-se que a variável p, que anteriormente, representava o nó pai, a partir deste momento passará a representar a probabilidade de transição entre estados.

A distribuição de probabilidade de transição de um estado para outro pode ser representado numa matriz de transição P[4] onde cada elemento da posição *(i,j)* representa a probabilidade de transição p_{ij}, do estado i para o estado j. Considerando a título ilustrativo, um caso em que existam três estados de transição possíveis, a matriz de transição será representada da seguinte forma:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}.$$

Na construção das cadeias de Markov utilizou-se a implementação do *package markovchain* [54].

4.5 *Multilayer Perceptrons*

As *Artificial Neural Network* (ANN), também conhecidas como *Neural Networks* (NN) são modelos computacionais designados à luz do funcionamento e estrutura dos neurónios humanos. Este modelo é constituído por um conjunto de unidades ou nós conectados, os *neurons*. As ANN, por conseguirem modelar complexas relações entre o *input* e o *output* são consideradas ferramentas estatísticas de modelação não linear. Existem dois tipos de ANN: as *feed-foward* e *recurrent networks*. Nessa dissertação serão utilizadas as MLP, que correspondem a redes *feed-foward*, para previsão do CLV.
Um *neuron* é constituído por 2 funções: a *net function* e a *activation fuction*, também conhecida por *transfer function*. A *net function* determina o modo como os sinais de *input*,x, são combinados dentro do *neuron*, usando para isso pesos w [5]. Assim, a *net fuction*, u, é dada por:

$$u = \sum_{i=0}^{N} x_i \times w_i \,. \tag{4.12}$$

O *output* de um *neuron* corresponde ao *output* da *net fuction* "transformada" pela função de ativação. Algumas das *neural transfer functions* mais utilizadas são: a tangente hiperbólica, a logística, a sigmóide, a *threshold*, a linear e a *staircase*. As funções de ativação devem ser diferenciáveis.
Numa MLP onde existe uma única *layer* apenas se consegue classificar dados linearmente separáveis.
A MLP é *"fully connected, feedforward neural network"*, isto significa que cada *layer* recebe como *input* o *output* de todos os *neurons* da *layer* imediatamente anterior, assim o neurónio da camada anterior contribuirá com o seu *output* para vários *neurons* da *layer* seguinte,

[4] Salienta-se que a variável P representava no capítulo 2, o valor de potencial do cliente, e a partir deste momento passará a representar a matriz de transição

[5] Salienta-se que a variável w que anteriormente representava as observações usadas como *input* a partir deste momento passará a representar os pesos. Sendo que w_0 é chamado de bias.

sendo por isso, necessário somar os erros dos *neurons* da camada da frente, propagados através dos respetivos coeficientes. Correspondendo, por isso, a um algoritmo de retro-propagação.

Para treinar uma MLP, é importante começar com pesos pequenos e aleatórios, próximos a 0, uma vez que a função de ativação sigmoidal pode saturar[6]. Também é importante executar o processo de treino, várias vezes, já que os dados de treino nem sempre são exatamente os mesmos. A normalização ou padronização do *input* é importante, pois caso as variáveis de *input* se encontrem em diferentes escalas forçarão a rede a ajustar pesos em taxas diferentes. Para treinar a rede, os dados de treino são apresentados com uma ordem aleatória. Uma passagem por todos os dados de treino corresponde a uma época. Em seguida, o processo é repetido até que o erro convirja ou seja detetado um ajuste excessivo. Salienta-se que durante o processo de treino, os pesos da *layer* de *output* e da *hidden layer* são otimizados.

Na construção da MLP utilizou-se a função *mlp* do *package* RSNNS [6], do R [41]. Após a criação da rede neuronal procedeu-se à sua parametrização, nesta etapa foram testadas várias alternativas para os seguintes parâmetros da função: *learning function*, *maxit*, *output*, *learning rate*, *size* e função de inicialização dos pesos.

Quando se constroem redes neurais é necessário fazer escolhas quanto aos valores iniciais dos pesos e bias. Se esta etapa for feita corretamente, a otimização será alcançada mais rapidamente, caso contrário, a convergência do gradiente descendente para o mínimo será mais demorada, ou até mesmo impossível. Os pesos e bias podem ser inicializados recorrendo à função *Randomize Weights*, que tal como o nome indica, inicializa todos os pesos[7] e bias[8] com valores que apresentam uma distribuição aleatória. Outra alternativa é a inicialização todos os pesos e bias com o mesmo valor, mas nesse caso, as *hidden layers* seriam todas iguais e por esse motivo haveria muito menos informação para conseguir chegar à solução ótima.

A função de aprendizagem (*Learning Function*) define a maneira como a aprendizagem ocorre na rede, sendo, por isso uma característica central da rede. No caso das MLP, existem várias funções de aprendizagem que podem ser utilizadas. Os algoritmos de *back-propagation* utilizam o gradiente descendente, ou seja, calcula-se o gradiente da função de erro em relação aos pesos da rede. O cálculo do gradiente inicia-se na última *layer* e prossegue pela rede até à primeira *layer*, sendo os cálculos parciais do gradiente de uma *layer* reutilizados no cálculo do gradiente da *layer* anterior. Nesta dissertação foram testadas 5 "variantes" de algoritmos de *backpropagation*: *Std Backpropagation*, *BackprpBatch*, *BackpropMomentum*, *BackpropChunk* e *BackpropWeightDecay*.

[6] Uma função de ativação é considerada saturada se estiver limitada a determinado intervalo, por exemplo o limite da função de ativação sigmóide quando tende para $-\infty$ é 0. O limite quando a função tende para $+\infty$ é 1. Assim, quando a saturação ocorre, o gradiente descendente tende a assumir valores muito pequenos, mesmo nos casos em que o erro de *output* é grande.

[7] Os pesos são valores que controlam a força da conexão entre dois neurónios, ou seja, os *inputs* são normalmente multiplicadas por pesos e isso define a influência que o *input* terá sobre o *output*.

[8] Correspondem a constantes que são adicionadas ao *input* ponderado antes de aplicar a função de ativação. O bias ajuda os modelos a representar padrões que não passam necessáriamnete pela origem.

A diferença entre a função *BackpropBatch* e a função *Std Backpropagation* está no momento em que a atualização dos pesos ocorre. Enquanto na função *Std Backpropagation* é executada uma etapa de atualização após cada época (passagem completa por todo o *training set*), na *BackpropBatch* todas as alterações de pesos são somadas e alteradas ao fim de um *batch* (conjunto de amostras processadas antes do modelo ser atualizado). No caso da função *BackpropWeightDecay*, após cada atualização, os pesos são multiplicados por um fator pouco menor que 1, impedindo dessa forma, crescimentos bruscos no valor dos pesos. A função *BackpropChunk* permite a atualização dos pesos em partes, bem como o treino seletivo dos *neurons*. Na função *BackpropMomentum* a introdução do *momentum rate* permite que haja uma atenuação das oscilações no gradiente descendente.
A função *Rprop* apenas utiliza o sinal da derivada para indicar a direção da atualização dos pesos.
A função *Quickprop* obtêm as direções de atualização através de minimizações unidimensionais e informações sobre a curvatura da função de erro.
A função *SCG* atualiza os valores de peso e bias de acordo com o método de gradiente conjugado em escala.
O *learning rate* (taxa de aprendizagem) é um hiperparâmetro que controla o quanto o modelo deve ser alterado em resposta ao erro estimado, de cada vez que os pesos do modelo são atualizados. Caso o valor do *learning rate* seja muito pequeno, o processo de treino será mais demorado, no entanto se o valor deste parâmetro for demasiado elevado também pode resultar num processo de treino instável ou na obtenção de conjunto de pesos não otimizados.
O parâmetro *maxit* define o número de época de treino a serem executadas, ou seja, quantas passagens completas são feitas pelo conjunto de dados. Se forem usadas poucas épocas, o modelo ficará propenso a problemas de *underfitting*, ou seja o modelo não aprende o suficiente para conseguir ter a capacidade de generalizar para novos conjuntos de dados. No entanto, se forem usadas muitas épocas, o modelo ficará propenso ao *overfitting*, ou seja, o modelo aprende demasiado bem com os dados do conjunto de treino, ajustando-se ao ruído do *dataset* inclusive e, perde capacidade preditiva quando é exposto a novos conjuntos de dados.
O parâmetro *size* define a quantidade de neurónios (*neurons*) presentes nas *hidden layers*.
A função de ativação do *output* pode ser definida com a opção *output ActFunc* ou com a opção *linOut*. A opção *linOut* permite definir a função de ativação do *output* como linear ou logística. Como a previsão do CLV é um problema de regressão foi definida como linear, ou seja, *linOut=TRUE*. A função de ativação logística é usada em tarefas de classificação.
A opção *outputActFunc* permite que a função de ativação definida seja utilizada por todas as unidades de *output*. Caso se opte por usar esta alternativa, teria de se utilizar a opção *outputActFunc = "Act Identity"*, uma vez que se pretende obter um *output* linear. Utilizando esta opção, estaria-se a definir a função "Act Identity"como sendo a função de ativação do *output* utilizada por todos os *neurons*.

4.6 Avaliação dos modelos supervisionados

O erro de treino obtém-se quando se executa o modelo nos mesmos dados que foram usados para o treino do modelo. O erro de teste obtém-se quando se executa o modelo treinado num conjunto de dados aos quais ele nunca foi exposto anteriormente.
Com o intuito de avaliar o erro de modo a que fosse facilmente interpretável em contexto empresarial, utilizaram-se 2 métricas: o erro médio absoluto (*Mean Absolute Error*, MAE) e o erro percentual.
O MAE resulta da média das diferenças, em valor absoluto, entre o valor previsto pelo modelo, $\hat{Y}$, para um cliente, i, e o seu valor de *target*, Y, ou seja,

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|. \tag{4.13}$$

O erro percentual expressa o MAE em percentagem do valor observado:

$$\text{Erro percentual} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i} \times 100. \tag{4.14}$$

4.7 *K-Means*

O *K-means* é um algoritmo de *clustering* não hierárquico.
O *clustering* é uma técnica de análise de dados que permite agrupar objetos semelhantes num grupo ou *cluster*, sendo os *clusters* diferentes entre si. Bons métodos de *clustering* produzem *clusters* com alta similaridade dentro do *cluster* e baixa similaridade entre *clusters*. O *clustering* difere dos problemas de classificação e regressão, porque no caso do *clustering* não há necessidade de conhecimento prévio das *labels* e, às vezes, não há uma noção clara daquilo que é *clustering* correto ou errado. Os algoritmos de *clustering* podem ser úteis quando não há conhecimento prévio sobre nenhum padrão, podendo também revelar novas informações sobre potenciais estruturas ocultas num determinado conjunto de dados.
O algoritmo *K-means* baseia-se na minimização da distância Euclidiana entre os centróides[9] e os membros associados a esses centróides.
Neste algoritmo, o número de *clusters* (k) é especificado *à priori*.
Considerando um conjunto de objetos, o *K-means* usa como método de agregação o método dos centróides. O algoritmo *K-means* começa inicializando k centróides. Inicialmente, os dados são atribuídos ao *cluster* mais próximo, de acordo com a distância Euclidiana aos centróides. Esta primeira etapa é designada por etapa de atribuição. De seguida, os centróides de cada *cluster* são atualizados iterativamente considerando todos os pontos de dados no *cluster*. Nesta etapa alguns objetos acabam por ser associados a outros *clusters*. O processo de associação dos *clusters* e de atualização dos centróides, é repetido até que não haja alterações no valor de nenhum dos centróides. Esta etapa é conhecida por

[9] Corresponde ao centro de cada *cluster* e é defenido pela média dos valores desse *cluster*.

convergência ou etapa da atualização.
Os resultados apresentados foram obtidos através da utilização da função *kmeans* do *package stats* [42] do R [41]. Um dos parâmetros de *input* deste método é o número de *clusters* sobre os quais se irá realizar o agrupamento, que nesta dissertação foi obtido pelo método do cotovelo. Este método calcula a distância de cada uma das observações até ao centróide do *cluster* ao qual ela pertence (soma dos quadrados intra-clusters). O ideal é que essa distância seja o menor possível. Neste método são testados diferentes números de *clusters*, e a soma dos quadrados intra-clusters diminui a cada incremento. Considera-se que se encontrou o número ideal de *clusters* quando o aumento no número de *clusters* não representa um valor significativo de ganho. Nesta dissertação, o gráfico do cotovelo foi obtido utilizando a métrica *withinss* da função *kmeans* do *package stats* [42] do R [41].
Geralmente, os algoritmos de *clustering* apresentam sensibilidade à alta dimensionalidade dos dados, uma vez que havendo muitas variáveis a tarefa de agrupamento fica dificultada [13]. Nesta dissertação, para reduzir o número de variáveis que seriam usadas como *inputs* no algoritmo *K-means* decidiu-se utilizar a função *importance* do *package randomForest* [34]. Esta função calcula a contribuição relativa de cada variável à tomada de decisão do algoritmo. De seguida, utilizou-se a função *cor* do *package stats* [42] para analisar a correlação que as variáveis apresentavam entre si.

Capítulo 5

Resultados

Neste capítulo serão apresentados os resultados da análise feita ao dataset incluindo os resultados da análise exploratória e descritiva e os resultados do ajuste dos modelos utilizados para prever o valor do CLV dos clientes. No final são apresentados os resultados decorrentes da análise de clusters.

5.1 Análise exploratória e descritiva do *dataset*

O *dataset* utilizado nesta dissertação apresentam uma grande heterogeneidade, uma vez que apenas 4 variáveis (Idade, Número de contas ativas, NPS *score* e NPS *score* do ano anterior) apresentam valores do coeficiente de variação inferiores a 50%.
A análise descritiva do *dataset* encontra-se subdividida em 4 tabelas: 5.1, 5.2, 5.3 e 5.4. Na tabela 5.1 são descritas as variáveis discretas, na tabela 5.2 as variáveis contínuas e na tabela 5.3 as variáveis "construídas" neste estudo. A análise descritiva das variáveis de rentabilidade é feita na tabela 5.4.
Nas secções 5.1.1 a 5.1.6 descrevem-se as variáveis estudadas, incluindo a representação gráfica das respetivas distribuições.

Tabela 5.1: Análise descritiva das variáveis discretas

Variável	**Min**	**1Q**	**$\tilde{x}$**	**$\overline{x}$**	**3Q**	**Max**	**DP**	**CV(%)**
Idade	18	40.0	53.0	53.4	67.0	109	17.7	33.1
Número de dependentes	0	0.0	0.0	0.2	0.0	9	0.6	284.1
Antiguidade	0	11.0	19.0	19.3	27.0	62	10.7	55.7
Número total de depósitos	0	0.0	0.0	0.4	0.0	42	1.1	283.3
Número de depósitos em ATM	0	0.0	0.0	0.1	0.0	32	0.8	501.2
Número de depósitos no balcão	0	0.0	0.0	0.2	0.0	41	0.8	316.3
Número total de transferências a crédito	0	0.0	2.0	2.6	4.0	231	3.7	145.8
Número total de compras a crédito	0	0.0	2.0	6.1	8.0	292	9.8	160.2
Número de *logins* no *site*	0	0.0	0.0	1.3	0.0	205	4.6	362.6
Número de contas ativas	1	1.0	1.0	1.1	1.0	11	0.4	36.6
Número total de depósitos no ano anterior	0	0.0	0.0	0.4	0.0	48	1.2	282.4
Número de depósitos em ATM no ano anterior	0	0.0	0.0	0.1	0.0	42	0.8	521.5
Número depósitos no balcão no ano anterior	0	0.0	0.0	0.3	0.0	48	0.8	313.2
Número total de transferências a crédito no ano anterior	0	0.0	2.0	2.2	4.0	391	3.0	137.3
Número total de compras a crédito no ano anterior	0	0.0	2.0	4.9	7.0	255	8.2	166.7
Número de *logins* no *site* no ano anterior	0	0.0	0.0	1.1	0.0	161	4.0	363.6
Número de idas à sucursal	0	0.0	0.0	0.2	0.0	19	0.7	310.2
Número de cartões de crédito *American Express*	0	0.0	0.0	0.2	0.0	16	0.6	262.0

Tabela 5.1: Análise descritiva das variáveis discretas

Variável	**Min**	**1Q**	**x̃**	$\overline{x}$	**3Q**	**Max**	**DP**	**CV(%)**
Número de cartões de crédito *Visa* e *Mastercard*	0	0.0	0.0	0.5	1.0	24	0.8	175.5
NPS score	0	7.7	8.5	8.0	9.0	9.0	1.4	17.7
Número de simulações	0	0.0	0.0	0.0	0.0	28	0.3	1202.6
NPS score no ano anterior	0	7.5	8.5	8.0	9.0	9.0	1.5	18.8
Número simulações no ano anterior	0	0.0	0.0	0.0	0.0	23	0.3	1095.1
Número de transações por iniciativa própria	0	4.0	14.0	19.6	28.0	339	19.8	101.3
Número de levantamentos	0	0.0	0.0	0.0	0.0	8	0.2	928.3
Número de outros depósitos	0	0.0	0.0	0.4	0.0	45	1.1	282.3
Número de relações familiares	0	0.0	1.0	0.9	1.0	8	1.0	114.1
Número de dias de resolução	0.0	0.0	0.0	0.7	0.0	22.0	3.0	428.6

Tabela 5.2: Análise descritiva das variáveis contínuas (em euros/mês)

Variável	**Min**	**1Q**	**x̃**	$\overline{x}$	**3Q**	**Max**	**DP**	**CV(%)**
Valor total de depósitos	5.0	160.0	350.0	468.3	650.0	1.9×10^3	405.3	86.6
Valor dos depósitos em ATM	5.0	120.0	280.0	375.5	550.0	1.4×10^3	318.3	84.8
Valor dos depósitos no balcão	10.0	164.6	350.0	487.2	864.9	2.1×10^3	438.3	90.0
Valor total de transferências a crédito	1.0	600.6	1.2×10^3	1.4×10^3	1.9×10^3	4.9×10^3	1.0×10^3	73.8
Valor total de compras a crédito	5.0	63.9	151.8	211.3	304.6	845.2	189.1	89.5
Valor total de depósitos no ano anterior	5.0	150.0	340.8	454.6	633.6	1.8×10^3	391.0	86.0
Valor dos depósitos em ATM no ano anterior	5.0	120.0	280.0	361.4	530.0	1.3×10^3	301.1	83.3
Valor dos depósitos no balcão no ano anterior	10.0	153.0	340.8	461.7	632.1	2.0×10^3	405.2	87.8
Valor total de transferências a crédito no ano anterior	1.0	606.8	1.2×10^3	1.4×10^3	1.9×10^3	4.7×10^3	1.0×10^3	72.7
Valor total de compras a crédito no ano anterior	5.0	54.8	129.3	182.9	262.8	741.3	165.3	90.4
Montante em produtos	1.0	1.0×10^3	6.5×10^3	2.0×10^4	2.8×10^4	1.1×10^5	2.6×10^4	135.5
Montante em produtos no ano anterior	1.0	734.0	4.8×10^3	1.6×10^4	2.2×10^4	1.1×10^5	2.4×10^4	145.4
Valor da poupança à habitação	20.4	48.7	417.0	417.0	815.3	815.3	560.8	134.5
Valor da poupança reforma	27.6	5.0×10^3	1.0×10^4	1.2×10^4	1.7×10^4	4.2×10^4	9.7×10^3	78.7
Valor das outras poupanças	5.6	2.0×10^3	6.5×10^3	1.1×10^4	1.6×10^4	5.5×10^4	1.2×10^4	109.7
Valor do crédito à habitação	121.0	2.5×10^4	4.7×10^4	5.4×10^4	10.0×10^4	1.9×10^5	3.8×10^4	70.3
Valor do crédito pessoal	26.0	1.5×10^3	4.7×10^3	4.2×10^3	1.3×10^4	2.4×10^4	3.8×10^3	89.7
Valor do seguro de vida	5.0	80.5	230.7	269.9	568.0	1.2×10^3	266.8	98.9
Valor do seguro automóvel	5.4	164.9	225.0	277.9	356.6	791.0	149.7	53.9

Tabela 5.2: Análise descritiva das variáveis contínuas (em euros/mês)

Variável	**Min**	**1Q**	$\tilde{\mathbf{x}}$	$\overline{x}$	**3Q**	**Max**	**DP**	**CV(%)**
Valor do seguro de saúde	7.6	279.2	519.9	684.2	1308.2	2.9×10^3	583.1	85.2
Valor do seguro de acidentes pessoais	1.0	36.9	66.6	78.9	133.2	258.9	51.9	65.7
Valor dos outros seguros	4.0	18.9	63.8	76.8	106.9	311.9	65.9	85.7
Valor dos recursos prazo	5.0	5.0	7.0	99.0	1000.0	2435.0	364.9	368.4
Valor do património financeiro	5.0	281.0	1.8×10^3	6.6×10^3	1.7×10^4	4.1×10^4	9.5×10^3	144.0
Valor dos recursos título	5.0	458.1	3.7×10^3	8.9×10^3	1.6×10^4	5.0×10^4	1.1×10^4	121.9
Valor dos recursos à ordem	5.0	191.9	663.8	1.7×10^3	4.2×10^3	1.0×10^4	2.3×10^3	132.2
Valor dos outros recursos	5.3	5.1×10^3	1.2×10^4	2.0×10^4	2.9×10^4	9.4×10^4	2.1×10^4	101.6
Valor do crédito vivo	25.0	244.8	1.2×10^3	1.4×10^4	4.7×10^4	1.1×10^5	2.5×10^4	172.6
Valor das simulações	291.0	5.0×10^3	1.3×10^4	3.5×10^4	1.0×10^5	2.1×10^5	4.6×10^4	131.4
Valor das simulações no ano anterior	200.0	5.0×10^3	1.4×10^4	3.9×10^4	5.4×10^4	2.3×10^5	5.0×10^4	130.7
Valor do saldo da conta corrente	0.0	318	1.1×10^3	6.1×10^3	4.4×10^3	1.1×10^4	2.6×10^4	417.9
Valor do seguro de multirriscos	5.2	86.5	126.4	141.9	180.4	387.6	75.5	53.2
Valor dos seguros de risco não vida	5.0	82.5	212.6	179.5	400.6	840.0	133.4	74.3
Valor estornado	0.7	5.4	6.2	11.5	10.8	147.3	16.0	139.1
Valor reclamado	0.5	5.6	15.8	276.0	21.0	3.0×10^3	235.9	907.7

Tabela 5.3: Análise descritiva das variáveis construídas nesta dissertação

Variável	Valores	Min	1Q	$\tilde{x}$	$\overline{x}$	3Q	Max	DP	CV(%)
Taxa de variação do valor total de depósitos	Positivos	0	0	0	5.8×10^{1}	0	3.2×10^{5}	1.9×10^{3}	3.2×10^{3}
	Negativos	-1.0×10^{2}	-1.0×10^{2}	-1.0×10^{2}	-7.9×10^{1}	-60	-2.5×10^{-3}	3.2×10^{1}	-4.0×10^{1}
Taxa de variação do número total de depósitos	Positivos	0	0	0	1.5×10^{1}	0	1.8×10^{3}	4.5×10^{1}	3.0×10^{2}
	Negativos	-1.0×10^{2}	-1.0×10^{2}	-1.0×10^{2}	-8.7×10^{1}	-80	-2.7	2.3×10^{1}	-2.6×10^{1}
Taxa de variação do número de depósitos no balcão	Positivos	0	0	0	1.1×10^{1}	0	1.2×10^{3}	3.5×10^{1}	3.3×10^{2}
	Negativos	-1.0×10^{2}	-1.0×10^{2}	-1.0×10^{2}	-9.1×10^{1}	-1.0×10^{2}	-5	2.0×10^{1}	-2.2×10^{1}
Taxa de variação do valor de depósitos no balcão	Positivos	0	0	0	4.0×10^{1}	0	2.8×10^{5}	1.5×10^{3}	3.7×10^{3}
	Negativos	-1.0×10^{2}	-1.0×10^{2}	-1.0×10^{2}	-8.3×10^{1}	-7.8×10^{1}	-2.5×10^{-3}	3.0×10^{1}	-3.6×10^{1}
Taxa de variação do valor total de transferências a crédito	Positivos	0	0	7.2	2.0×10^{2}	100	1.3×10^{6}	5.1×10^{3}	2.5×10^{3}
	Negativos	-1.0×10^{2}	-6.8×10^{1}	-1.4×10^{1}	-3.4×10^{1}	-2.1	-3.2×10^{-4}	3.8×10^{1}	-1.1×10^{2}
Taxa de variação do número total de transferências a crédito	Positiva	0	0	0	39.88655	100	3.5×10^{4}	1.2×10^{2}	3.0×10^{2}
	Negativa	-1.0×10^{2}	-1.0×10^{2}	-50	-6.2×10^{1}	-3.6×10^{1}	-1.1	2.9×10^{1}	-4.7×10^{1}
Taxa de variação do número total de compras a crédito	Positiva	0	0	1.5×10^{1}	8.5×10^{1}	1.0×10^{2}	6.9×10^{3}	2.0×10^{2}	2.4×10^{2}
	Negativa	-100	-100	-5.7×10^{1}	-6.0×10^{1}	-3.3×10^{1}	-1.4	3.2×10^{1}	-5.3×10^{1}
Taxa de variação do número de *logins* no *site*	Positiva	0	0	0	2.0×10^{1}	0	8.2×10^{3}	9.2×10^{1}	4.5×10^{2}
	Negativa	-1.0×10^{2}	-1.0×10^{2}	-80	-7.2×10^{1}	-50	-1.8	3.0×10^{1}	-4.3×10^{1}
Taxa de variação da rentabilidade líquida dos últimos 12 meses	Positiva	0	1.2×10^{1}	3.4×10^{1}	3.9×10^{2}	1.0×10^{2}	23.9×10^{5}	1.1×10^{4}	2.8×10^{3}
	Negativa	-1.0×10^{2}	-5.9×10^{1}	-3.1×10^{1}	-3.2×10^{1}	-1.5×10^{1}	-2.2×10^{-3}	2.3×10^{1}	-7.4×10^{-1}

Tabela 5.3: Análise descritiva das variáveis construídas nesta dissertação

Variável	**Valores**	**Min**	**1Q**	$\tilde{\mathbf{x}}$	$\overline{x}$	**3Q**	**Max**	**DP**	**CV(%)**
Taxa de variação do montante de produtos	Positiva	0	8.6	3.3×10^{1}	1.0×10^{4}	1.3×10^{2}	3.7×10^{7}	3.0×10^{5}	3.0×10^{3}
	Negativa	-100	-3.8×10^{1}	-1.4×10^{1}	-2.5×10^{1}	-4.5	-7.7×10^{-5}	2.6×10^{1}	-1.0×10^{2}
Taxa de variação do valor das simulações	Positiva	0	0	0	1.6	0	7.2×10^{3}	2.9×10^{1}	1.8×10^{3}
	Negativa	-100	-100	-1.0×10^{2}	-9.8×10^{1}	-1.0×10^{2}	-6.4×10^{-1}	9.5	-9.7
Taxa de variação do número de simulações	Positiva	0	0	0	1.4	0	2.6×10^{3}	1.5×10^{1}	1.1×10^{3}
	Negativa	-1.0×10^{2}	-1.0×10^{2}	-1.0×10^{2}	-9.9×10^{1}	-1.0×10^{2}	-25	7.1	-7.2

Tabela 5.4: Análise descritiva das variáveis de rentabilidade

	Rentabilidade líquida dos últimos 12 meses		**Rentabilidade 2 anos antes**		**Target (CLV)**	
	Positiva	Negativa	Positiva	Negativa	Positiva	Negativa
Min	0	-353.2	0	-351.4	0	-338.2
Q1	54.8	-80.5	56.5	-93.6	60.1	-71.5
$\tilde{\mathbf{x}}$	98.2	-18.9	104.5	-16.7	99.6	-21.3
$\bar{x}$	160.6	-61.0	161.5	-65.7	166.4	-58.1
Q3	217.2	-6.6	219.6	-7.5	222.7	-7.0
Máx	723.9	-1×10^{-2}	700.5	-1×10^{-2}	715.6	-1×10^{-2}
DP	155.6	85.5	151.6	91.0	159.8	80.0
CV	96.9	-140.2	93.9	-138.5	96.0	-137.7

5.1.1 Variáveis demográficas

5.1.1.1 Grau de instrução

O estudo do grau de instrução foi feito considerando diferentes códigos de grau de instrução, que indicam o nível de escolaridade completa que o cliente possui (ver tabela 5.5). Concluiu-se que cerca de 40% dos clientes têm o 3º ciclo completo. Segue-se a classe dos clientes que têm o ensino superior e o 1º ciclo com percentagens próximas (cerca de 17% e 15%, respetivamente). A classe em que se encontram menos clientes é na dos cursos técnicos (cerca de 2%).
A análise do valor do CLV em função da grau de instrução, apresentada na figura 5.1, indica que os clientes que concluiram o ensino universitário são aqueles que apresentam uma maior distância interquartil. Pelo contrário, os clientes sem estudos são aqueles que apresentam uma menor distância interquartil. Os *boxplots* dos clientes que tem o 2º ciclo ou o ensino secundário são idênticos, sendo que 75% dos clientes tem um CLV inferior a (cerca de) 250 euros. Comparando o *boxplot* destes clientes com o *boxplot* dos clientes que tem um curso técnico, verifica-se que o valor do 3º quartil é ligeiramente superior para os clientes que possuem um curso técnico.

Tabela 5.5: Distribuição dos clientes por grau de instrução

Código	**Descrição**	**%**
A	3º Ciclo	40.3
B	Ensino Universitário	16.6
C	1º Ciclo	15.2
D	2º Ciclo	12.6
E	Ensino secundário	9.9
F	Sem estudos	2.9
G	Curso técnicos	2.4

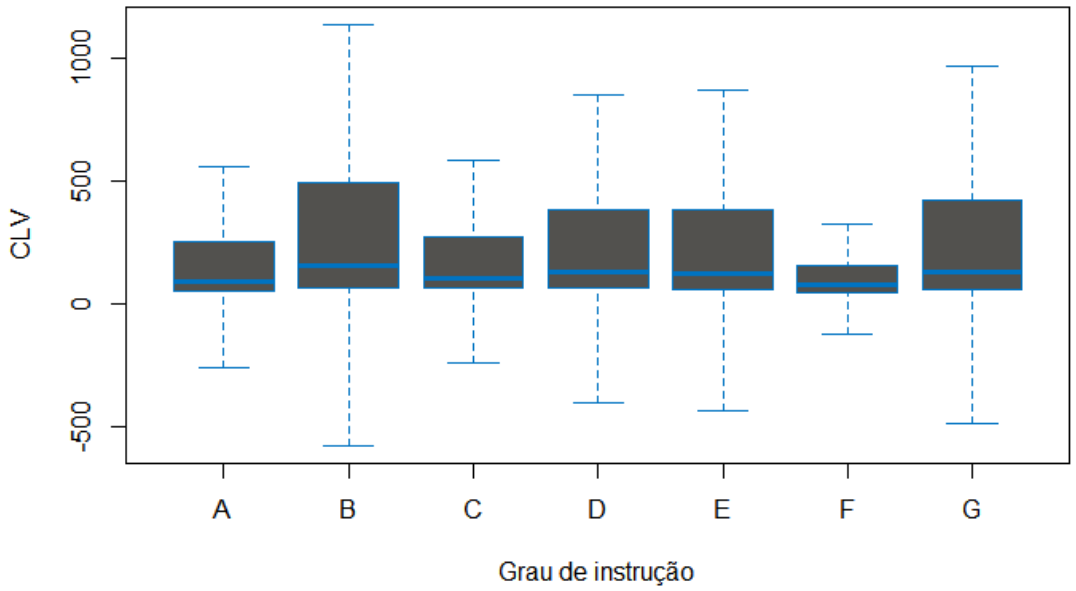


Figura 5.1: Variação do CLV (em euros) em função do grau de instrução (ver legenda na tabela 5.5)

5.1.1.2 Profissão

Na tabela 5.6 encontra-se a descrição de cada um dos códigos de profissão utilizados nesta análise bem como, a percentagem de clientes existentes em cada profissão. Salienta-se, que a maior parte dos cliente são trabalhadores não qualificados, seguindo-se os profissionais das forças armadas.

Tabela 5.6: Distribuição dos clientes por profissão

Descrição	**%**
Trabalhadores não qualificados	37.0
Profissões das forças armadas	12.9
Profissões de prestação de serviços	10.3
Especialistas das atividades intelectuais e cientificas	8.8
Representantes do poder legislativo e órgãos	7.5
Trabalhadores qualificados da indústria	7.5
Administrativos	7.0
Operadores de instalações e máquinas	4.0
Técnicos e profissões de nível intermédios	3.9
Agricultores e trabalhadores qualificados	3.9

5.1.1.3 Situação Profissional

Na tabela 5.7 resume-se-se a distribuição dos clientes por profissão. Salienta-se, que cerca de metade dos clientes trabalha por conta de outrém. Os reformados são o grupo que aparecem em maior percentagem a seguir aos trabalhadores por conta de outrém, podendo este, ser um indicador de envelhecimento a ser tido em conta.

Tabela 5.7: Distribuição dos clientes por situação profissional

Descrição	**%**
Trabalhadores por conta de outrém	53.9
Reformados	17.3
Desempregado	13.1
Estudante	7.8
Trabalhadores por conta própria	5.6
Outros	2.3

5.1.1.4 Sexo

A distribuição dos clientes por sexo está equilibrada, existindo, no entanto, uma percentagem ligeiramente maior de homens na amostra (57% de homens *vs* 43% de mulheres). Na figura 5.2 apresenta-se a distribuição do CLV em função do sexo, podendo-se concluir que, os valores do CLV são mais heterogéneos no grupo masculino, embora a mediana seja aproximadamente a mesma nos 2 grupos.

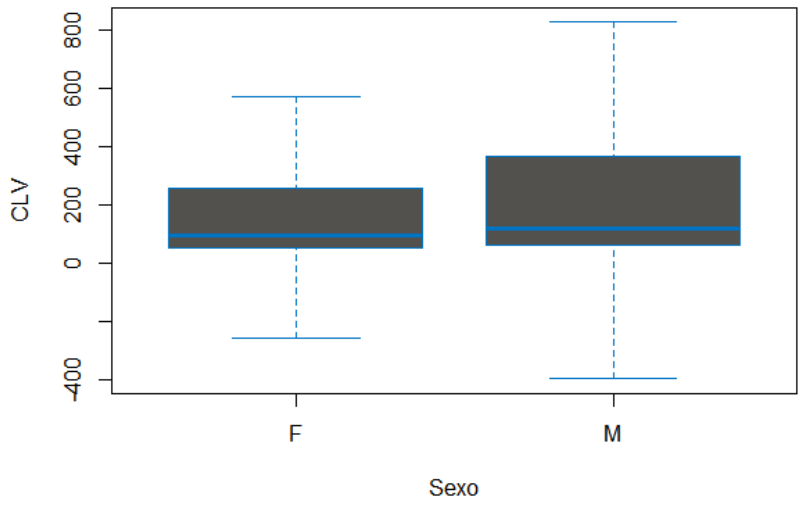


Figura 5.2: Distribuição do CLV(euros/mês) em função do sexo. (F - Feminino; M-Masculino).

5.1.1.5 Estado civil

Nesta amostra predominam os estados civis casado ou solteiro, sendo cada um dos grupos constituído por cerca de 50% e 40% da amostra, respectivamente. O terceiro grupo com mais indivíduos é o dos divorciados que, contém 7.8% da amostra (ver tabela 5.8). Quanto à análise do CLV em função do estado civil, apresentada na figura 5.3, verifica-se que os clientes que vivem em união de facto são os que apresentam uma maior distância interquartis. Neste grupo 75% dos clientes apresentam uma rentabilidade inferior ou igual a cerca de 500 euros. O grupo dos viúvos é o que apresenta uma distância interquartil menor.
O primeiro quartil e a mediana encontram-se relativamente próximos em todos os grupos.

Tabela 5.8: Distribuição dos clientes por estado civil

Código	Descrição	%
A	Casado	49.6
B	Solteiro	37.1
C	Divorciado	7.8
D	Viúvo	4.3
E	União de facto	1.1

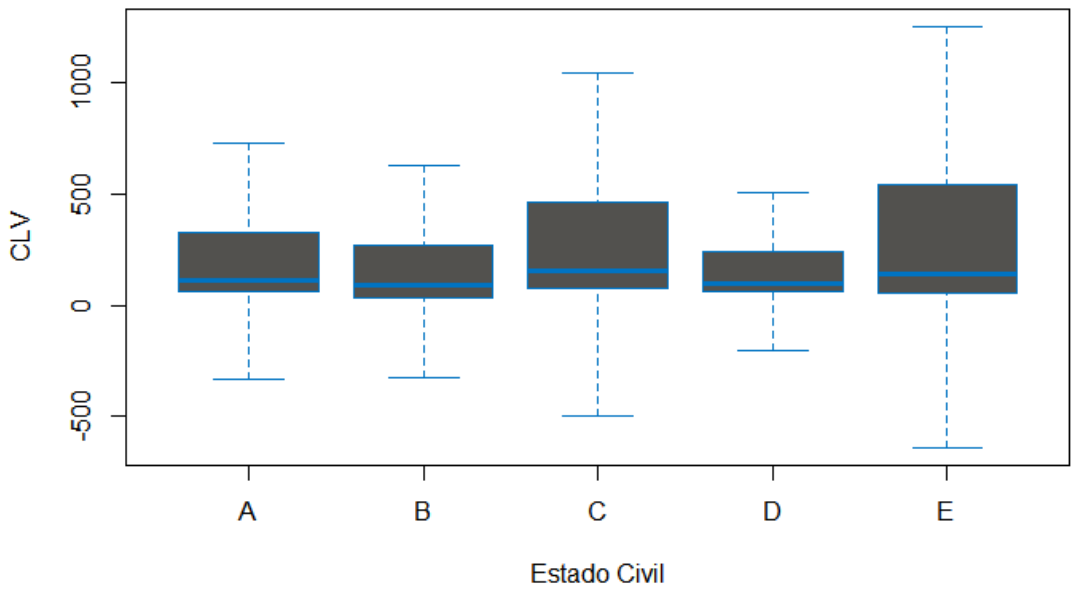


Figura 5.3: Variação do CLV (em euros/mês) em função do estado civil (ver legenda na tabela 5.8)

5.1.1.6 Número de dependentes

Os clientes que não tem dependentes encontram-se em maioria neste *dataset*, representando aproximadamente 88% dos clientes da amostra. Seguem-se os que tem um dependente (6.8%) e dois dependentes (4.4%). Salienta-se que apenas 1.1% dos clientes tem 3 ou mais dependentes a seu cargo (ver tabela 5.9).

Tabela 5.9: Distribuição por número de dependentes

Número de dependentes	%
0	87.8
1	6.8
2	4.4
3	0.8
4	0.2
≥5	0.1

5.1.1.7 Número de relações familiares

Esta variável, contabiliza o número de familiares (relações parentais e matrimoniais) do cliente que também têm conta no banco. Pela análise da tabela 5.10 verifica-se que 78.7% dos clientes ou não tem nenhum, ou têm apenas um familiar com conta no banco. Apenas cerca de 2% tem um número de relações familiares superior ou igual a 4.

Tabela 5.10: Distribuição dos clientes por número de relações familiares

Número de relações familiares	%
0	44.2
1	34.5
2	14.2
3	5.2
≥ 4	1.8

5.1.2 Variáveis transaccionais

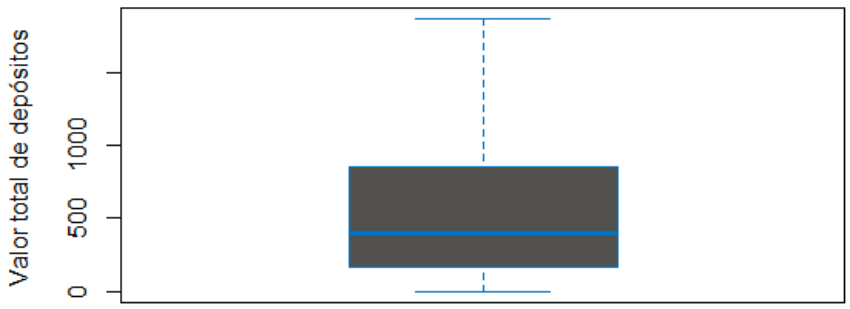
5.1.2.1 Depósitos

Na análise dos depósitos considerou-se o número de depósitos e o seu montante. Os depósitos podem ser feitos em 2 tipos de canais diferentes: ATM ou balcão. Esta distinção por canal foi tida em conta na análise. Na tabela 5.11 encontra-se sumariado o número total de depósitos efetuados bem como o número depósitos por canal.
Nas figuras 5.4, 5.5 e 5.6 encontra-se representada a distribuição do valor total de depósitos, do valor dos depósitos em ATM e do valor do depósito no balcão, respetivamente. Os 3 gráficos mostram uma distribuição assimétrica positiva, com claro predomínio de

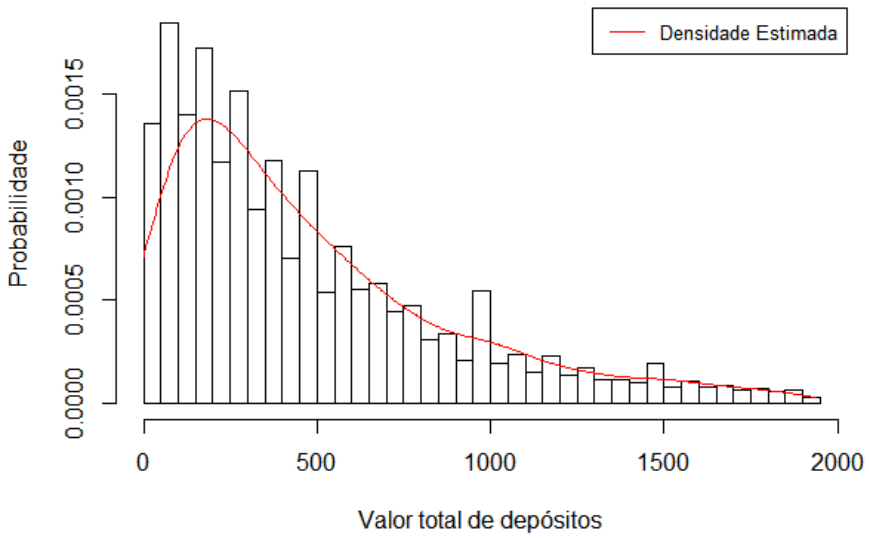
valores depositados inferiores a 500 €. Verifica-se que quando os clientes pretendem fazer depósitos de maior valor o canal escolhido é o balcão: 75% dos depósitos feitos em ATM apresentam valores inferiores ou iguais (a cerca de) 700 €, enquanto que no caso dos valores depósitos feitos em balcão verificou-se que 75% apresentava valores inferiores ou iguais (a cerca de) 1000 €.

Tabela 5.11: Distribuição da percentagem de depósitos por canal

Número de depósitos	**% de depósitos em ATM**	**% de depósitos no balcão**	**% total de depósitos**
0	92.2	83.2	77.2
1	4.6	12.0	14.5
2	1.7	3.0	4.6
3	0.7	0.9	1.8
≥ 4	0.8	0.8	2.0

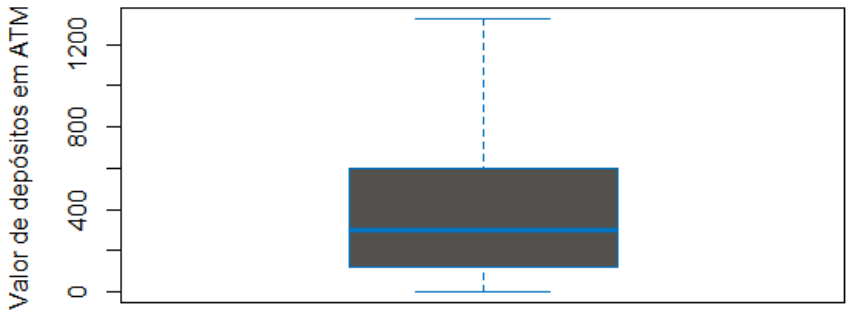


(a)

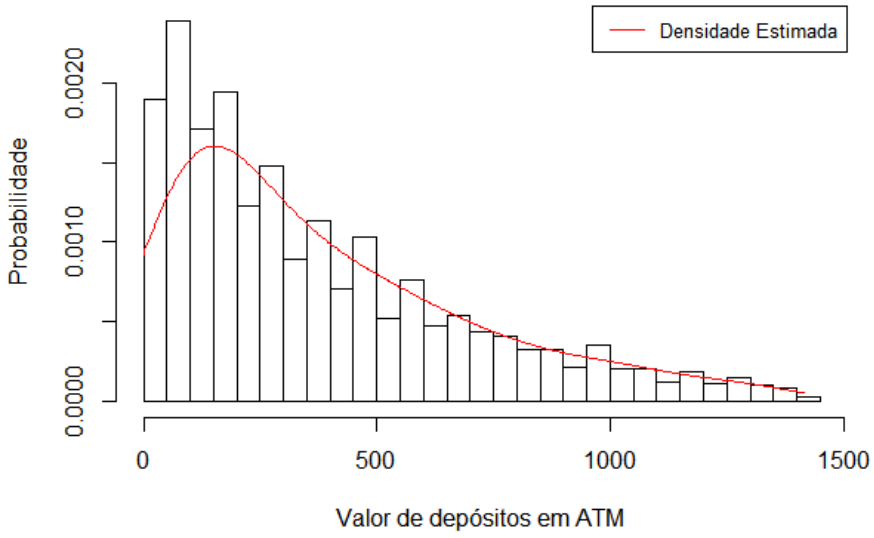


(b)

Figura 5.4: Distribuição do valor total de depósitos (euros/mês)

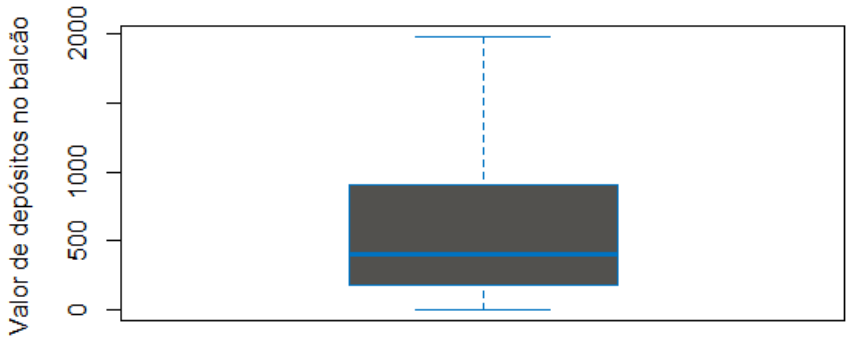


(a)

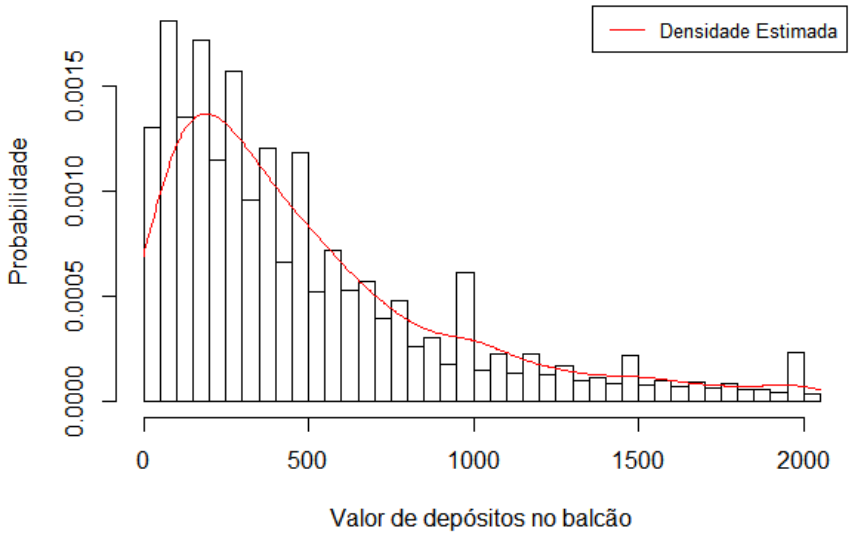


(b)

Figura 5.5: Distribuição do valor de depósitos em ATM (euros/mês)



(a)



(b)

Figura 5.6: Distribuição do valor do depósitos no balcão depósitos (euros/mês).

5.1.2.2 Compras a crédito

Nas figuras 5.7[1] e 5.8 analisa-se o número de compras a crédito feitas por cliente e o valor que despendem nas mesmas, respetivamente. Cerca de 35% dos clientes da amostra não fez qualquer compra a crédito no mês considerado nesta análises. Verifica-se que os clientes desta amostra não tem por hábito fazer muitas de compras a crédito por mês, uma vez que, a percentagem de clientes que fez 20 ou mais compras a crédito é pouco expressiva.
A mediana do valor das compras a crédito ronda os 200 €, apesar de os valores mais frequentes das compras a crédito variarem entre os 20 € e os 60 €.

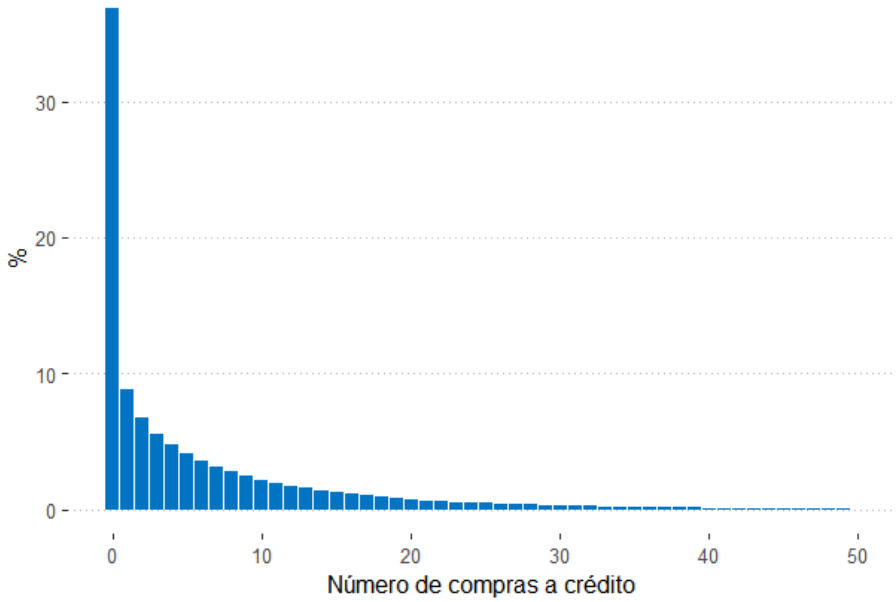
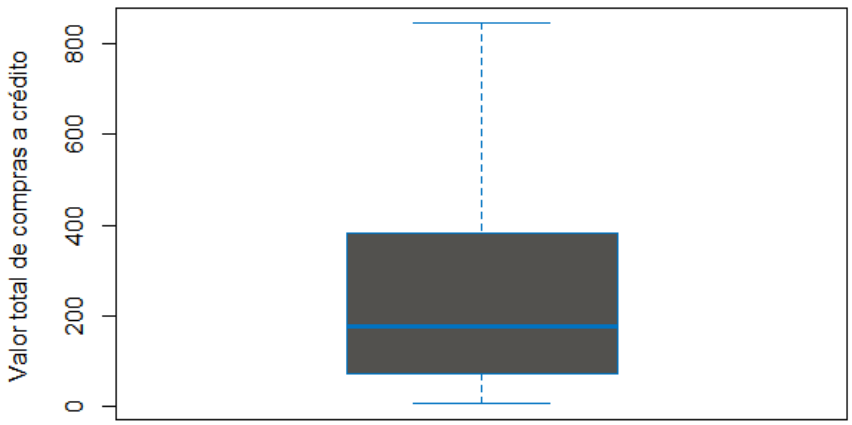
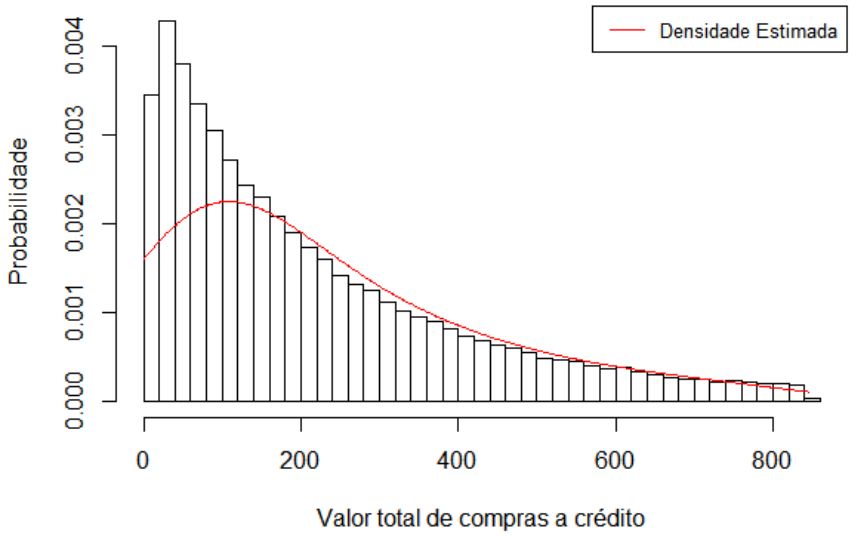


Figura 5.7: Distribuição do número de compras a crédito.

[1] O gráfico apresenta apenas os casos em que o número total de compras a crédito efetuadas por mês é inferior a 50, o que corresponde a 99.3% da amostra.



(a)



(b)

Figura 5.8: Distribuição dos valores de compras a crédito (euros/mês)

5.1.2.3 Transferências a crédito

Na figura 5.9[2] encontram-se representada a percentagem do número de transferências a crédito efetuadas mensalmente.
Da análise do *boxplot*, na figura 5.10, verifica-se que a distância entre o primeiro quartil e a mediana é idêntica à distância entre a mediana e o terceiro quartil, o que significa que, a distribuição desta variável é "aproximadamente" simétrica. Analisando o gráfico de densidade, verifica-se uma menor tendência das transferências a crédito com valores superiores a 1500 €.

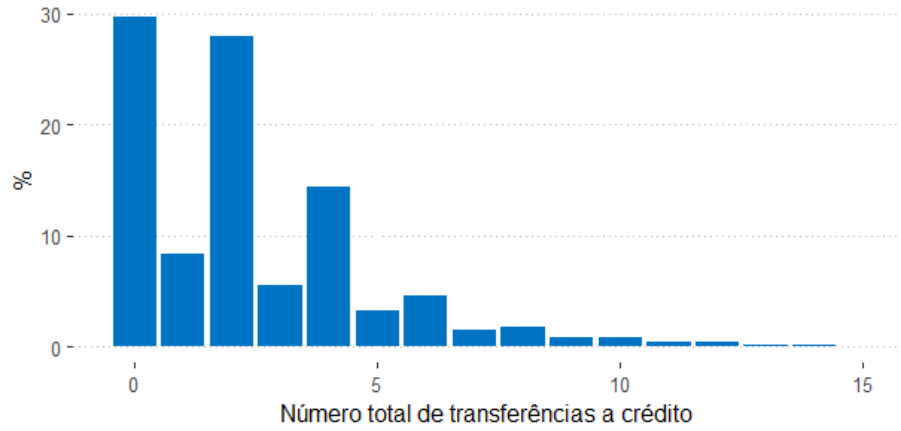
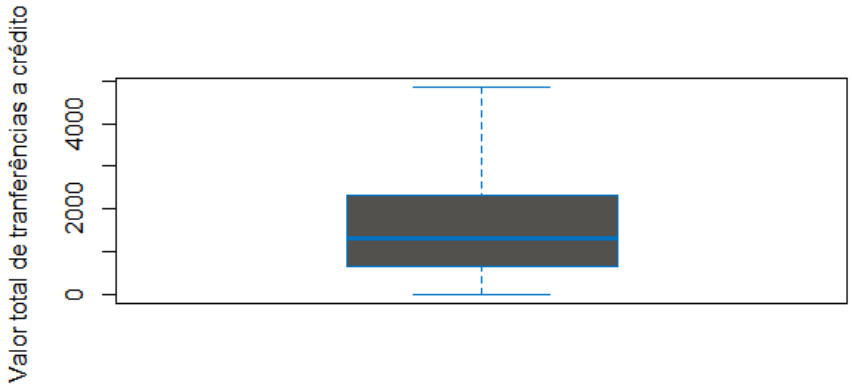
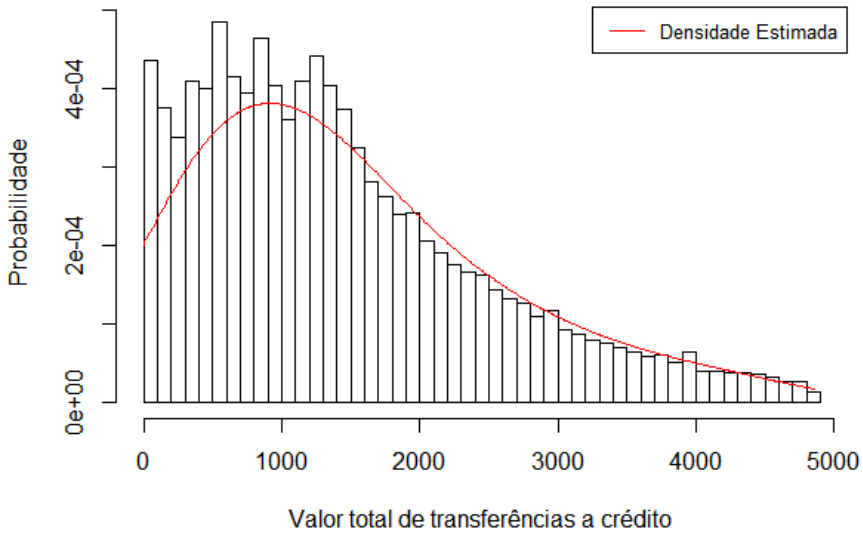


Figura 5.9: Distribuição do número de transferências a crédito (euro/mês).

[2]O gráfico apresenta apenas os casos em que o número total de transferências a crédito efetuadas por mês é inferior a 15, o que corresponde a 99.1% da amostra.



(a)



(b)

Figura 5.10: Distribuição do valor das transferências a crédito (euros/mês)

5.1.2.4 Número de levantamentos

Nos últimos anos, tem-se verificado uma tendência decrescente na utilização de numerário. Através da tabela 5.12 é perceptível que essa tendência também existe nos clientes deste banco, uma vez que, a maioria dos clientes não fez qualquer levantamento de dinheiro.

Tabela 5.12: Número de levantamentos (por mês)

Número de levantamentos	%
0	98.4
1	1.3
≥ 2	0.03

5.1.2.5 Transações por iniciativa própria

O número de transações por iniciativa própria, apresentadas na figura 5.11[3] é bastante heterogéneo. Apesar de a maioria dos clientes fazer uma, duas ou nenhuma transação, também há clientes que fazem mais de 100 transações por sua iniciativa.
Nota: Pagamentos "obrigatórios", como por exemplo o pagamento da água ou luz, não foram consideradas como transações por iniciativa própria do cliente.

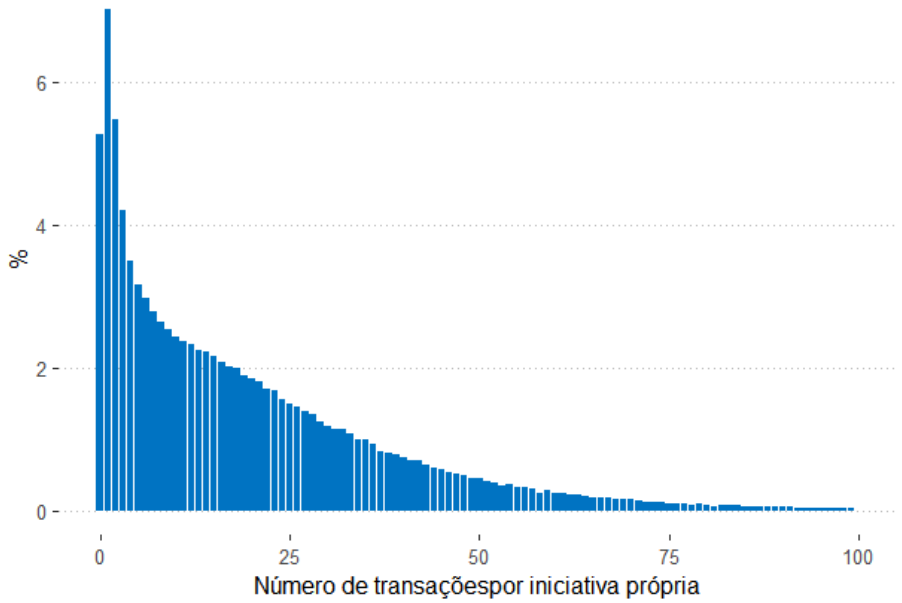


Figura 5.11: Distribuição do número de transações por iniciativa própria (por mês).

[3]No gráfico são apresentados apenas os casos em que o número total de transações por iniciativa própria efetuadas por mês é inferior a 100, o que corresponde a 99.5% da amostra.

5.1.3 Variáveis de fidelização

5.1.3.1 Idas a sucursal

A maioria dos clientes não se desloca à sucursal. Apenas 1.4% dos clientes foram mais de três vezes à sucursal durante o mês analisado (ver tabela 5.13).

Tabela 5.13: Número de idas à sucursal (por mês)

Número de idas à sucursal	**%**
0	86.8
1	9.3
2	2.5
≥ 3	1.4

5.1.3.2 *Logins* ao site

A grande maioria dos clientes (79.6%) não fez nenhum *login* no site, verificando-se também, um decréscimo no número de clientes à medida que se considera um maior número de *logins* (ver tabela 5.14).

Tabela 5.14: *Logins* site (por mês)

***Logins* site**	**%**
0	79.6
1	5.5
2	3.4
3	2.3
4	1.7
≥ 5	7.5

5.1.3.3 Número de contas ativas

Por norma os clientes possuem apenas uma conta no banco, no entanto cerca de 11% dos clientes possui 2 contas. É raro haver clientes que tenham 3 ou mais contas (ver tabela 5.15).

Tabela 5.15: Contas ativas

Número de contas ativas	**%**
1	87.5
2	10.8
≥ 3	1.7

5.1.3.4 NPS *score*

O *NPS score* é uma variável discreta e ordinal, que varia de 0 até 9 e mede o grau de satisfação do cliente com a seu banco. Para interpretar a variação desta variável foram considerados 3 intervalos: se o cliente atribuir uma classificação entre 0 e 3, significa que se encontra pouco satisfeito com o banco; caso atribua uma classificação entre 4 e 6, significa que está satisfeito; se o valor atribuído for superior a 6, então o cliente está muito satisfeito com o seu banco. De acordo com a tabela 5.16 conclui-se que 89.9% dos clientes encontra-se muito satisfeito com o banco.

Tabela 5.16: NPS score

NPS *score*	**%**
[0,3]	1.7
]3,6]	8.4
]6,9]	89.9

5.1.3.5 Antiguidade no banco

Em relação à antiguidade do cliente no banco verifica-se que, apesar de haver clientes que começaram a relação com o banco recentemente também há outros que são clientes do banco há longos anos (figura 5.12[4]).

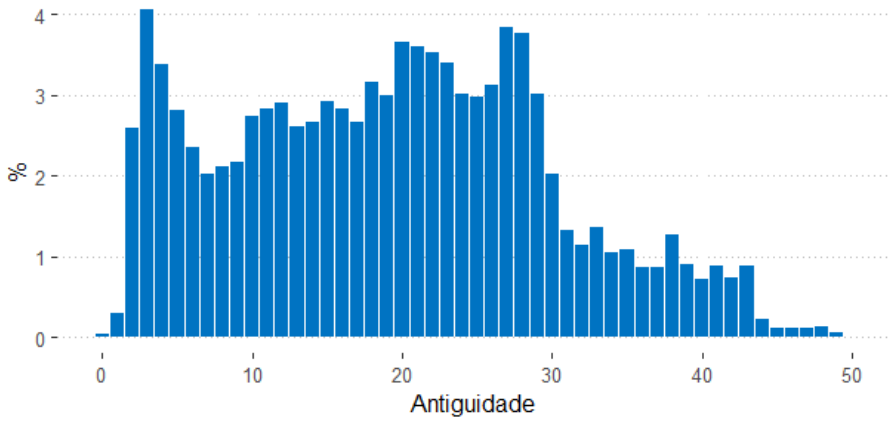
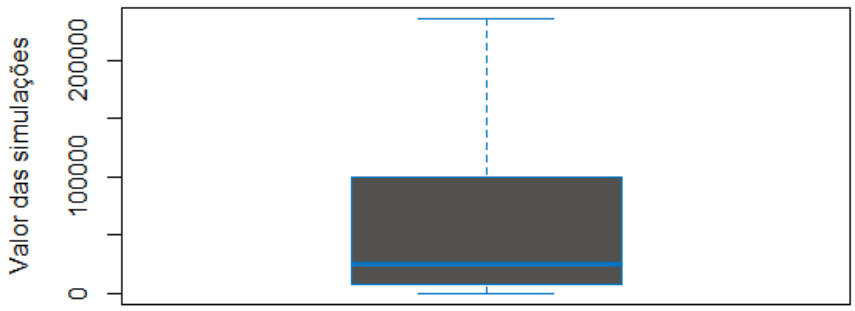


Figura 5.12: Antiguidade dos clientes (anos).

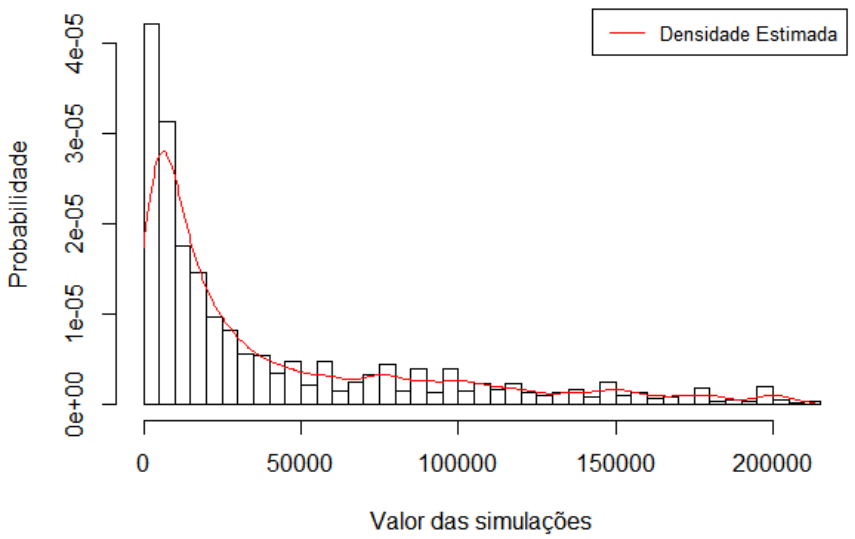
[4]Salienta-se que, apenas 4 clientes tem uma antiguidade no banco superior a 50 anos. Esses clientes não se encontram representados na figura 5.12.

5.1.3.6 Valor das simulações

Na variável "valor das simulações" não foi tido em conta o produto para o qual o cliente fez a simulação, por isso, existe uma grande amplitude de valores para esta variável. Cerca de 75% dos clientes fez simulações para valores inferiores ou iguais a 100000 €. Esta variável apresenta uma assimetria positiva.



(a)



(b)

Figura 5.13: Distribuição do valor das simulações (euros/mês).

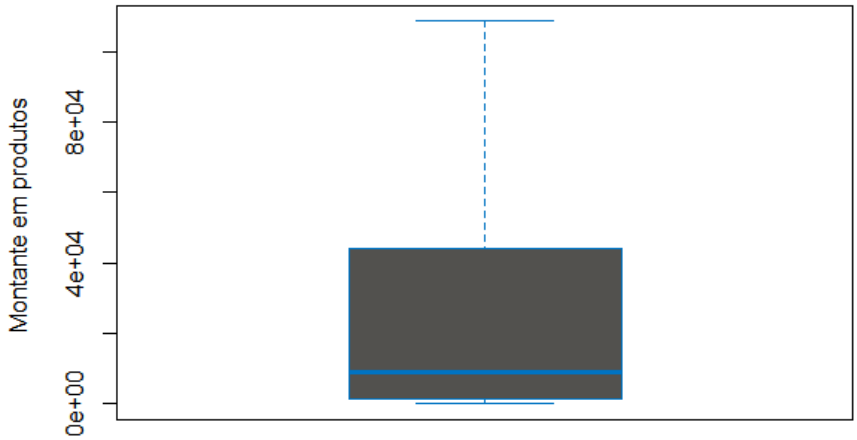
5.1.3.7 Clientes ativos

Nesta amostra foram considerados como clientes ativos aqueles que registaram pelo menos uma transação nos últimos 3 meses ou tem um saldo de conta corrente superior a 100 €. Verificou-se que 88.1% dos clientes encontram-se ativos.

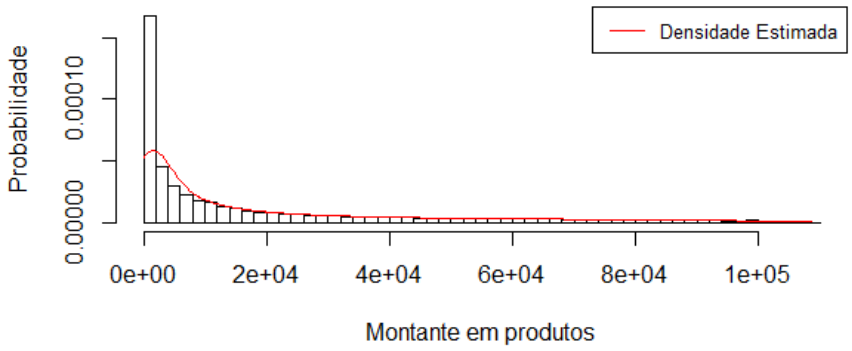
5.1.4 Variáveis de posse

5.1.4.1 Montante em produtos (*cross-sel*)

Na figura 5.14 encontra-se representado o *boxplot* e a distribuição do montante resultante da posse de produtos *cross-selling*. Quanto ao *boxplot* verifica-se que a distância interquartil entre o segundo e o terceiro quartil é superior relativamente às restantes distâncias interquartis. O primeiro quartil e a mediana apresentam valores relativamente próximos. Esta variável apresenta uma distribuição assimétrica positiva (ver figura 5.14). Os clientes apresentam, mais frequentemente, montantes em produtos *cross- selling* inferiores a 2000€.



(a)

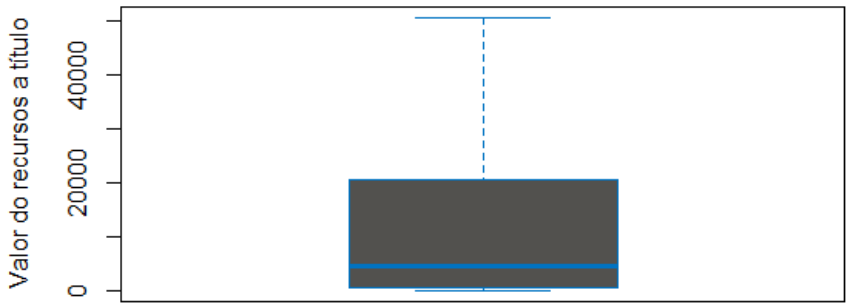


(b)

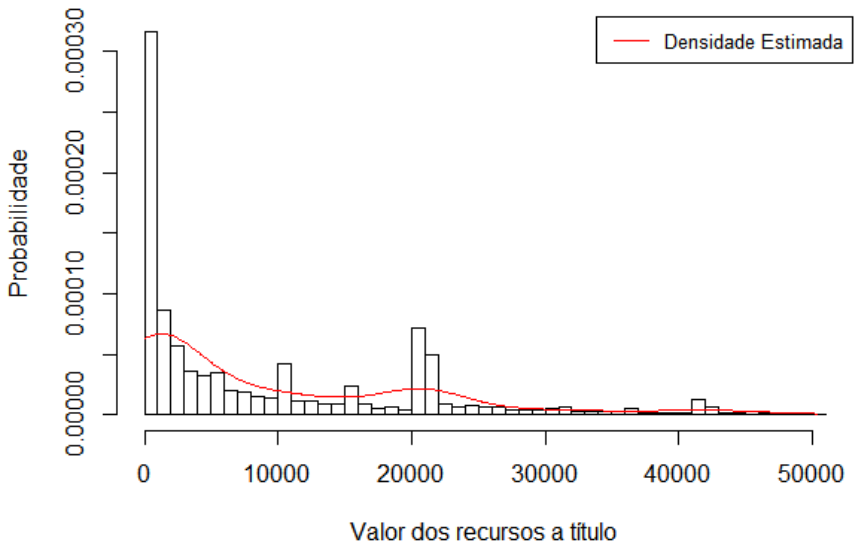
Figura 5.14: Distribuição do montante em produtos (euros/mês).

5.1.4.2 Recursos a título

Analisando a figura 5.15, concluí-se que metade dos clientes possuem recursos a títulos com valores inferiores a (cerca de) 4000€. Verifica-se também que, os valores mínimos se encontram muito próximos do 1º quartil. Apesar de o gráfico de distribuição desta variável apresentar uma tendência decrescente, verifica-se também algumas oscilações, por exemplo, os recursos a título com valores entre 20000 e 22000€ apresentam uma maior frequência relativamente aos valores de recursos a título antecedentes e precedentes.



(a)

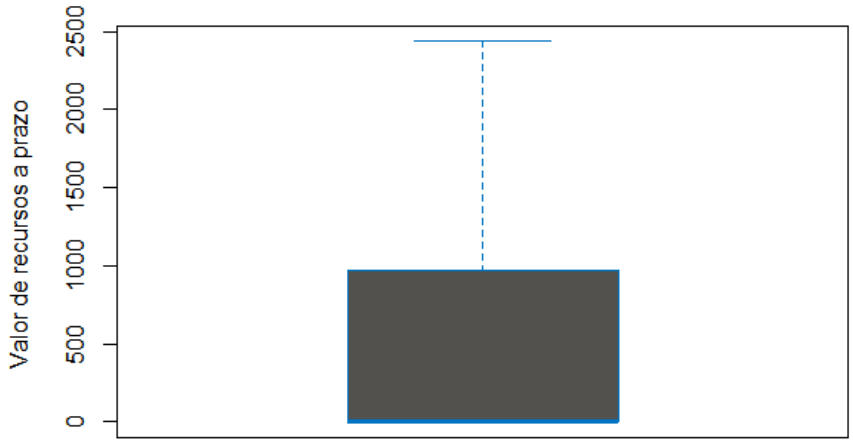


(b)

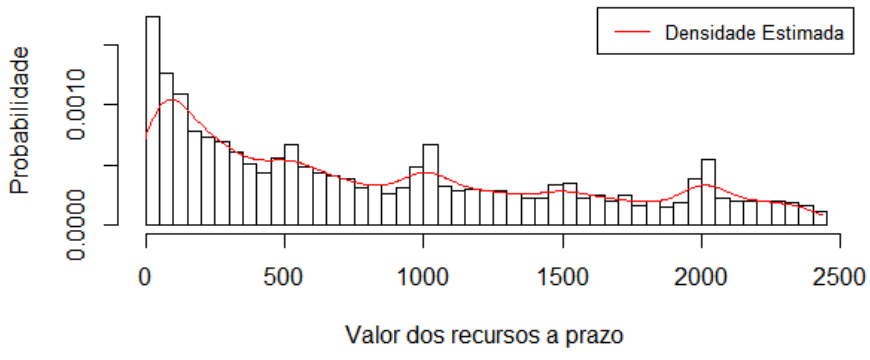
Figura 5.15: Distribuição do valor dos recursos a título (euros/mês).

5.1.4.3 Recursos a prazo

Em relação ao valor dos recursos a prazo verifica-se que o primeiro e segundo quartil apresentam valores muito próximos. Apesar dos recursos a prazo com valores inferiores a 500€ serem os mais frequentes, verifica-se que os recursos com valores de 1050€ e 2050€ são mais frequentes do que os valores que lhe antecedem e precedem (ver figura 5.16).



(a)

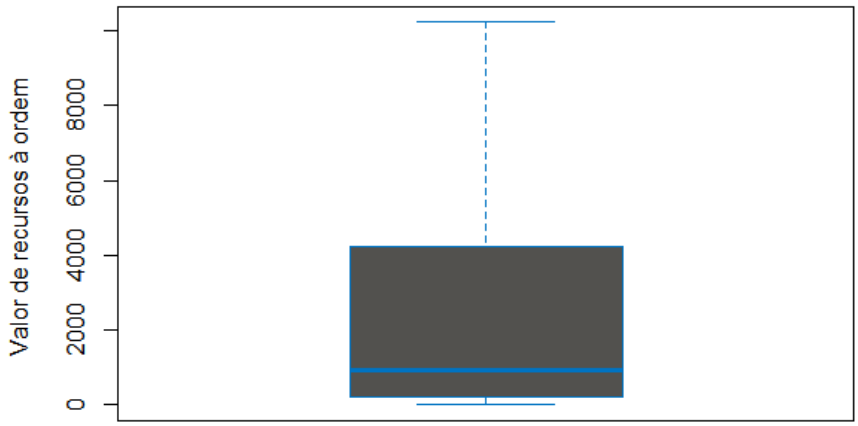


(b)

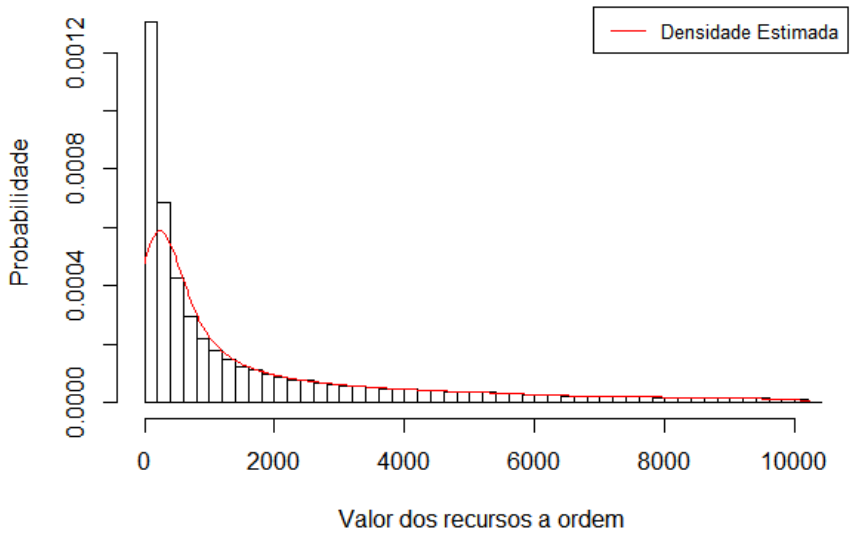
Figura 5.16: Distribuição do valor dos recursos a prazo (euros/mês).

5.1.4.4 Recursos à ordem

No *boxplot* do valor de recursos à ordem, verificou-se que, a amplitude entre o segundo e o terceiro quartil é superior à distância interquartis existente entre o primeiro e o segundo quartil. Esta variável tem uma distribuição assimétrica positiva, em que à medida que o valor dos recursos à ordem aumenta, o número de clientes diminui.



(a)

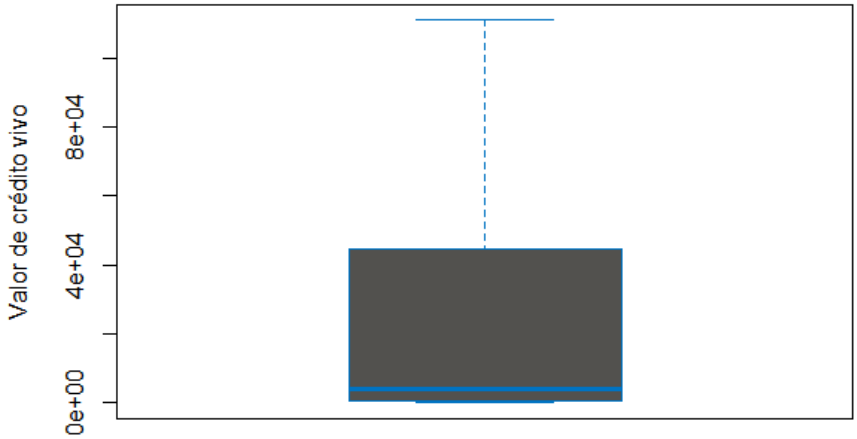


(b)

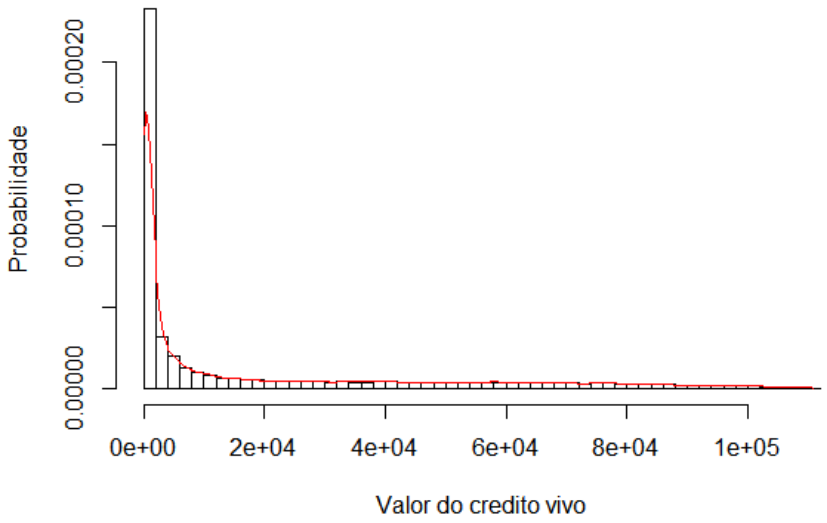
Figura 5.17: Distribuição do valor de recursos à ordem (euros/mês)

5.1.4.5 Valor de crédito vivo

A variável crédito vivo apresenta uma grande amplitude de valores que, poderá dever-se ao facto não ser feita referência ao tipo de crédito subjacente a cada um dos casos. O valor mínimo desta variável situa-se próximo dos 0 €. O máximo situa-se na ordem dos 10^5 € (ver figura 5.18).



(a)

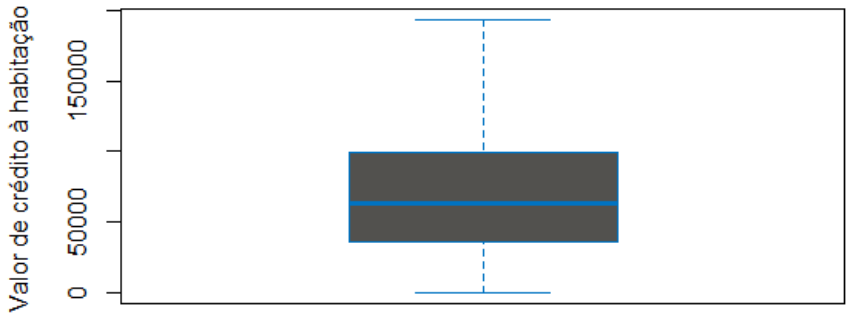


(b)

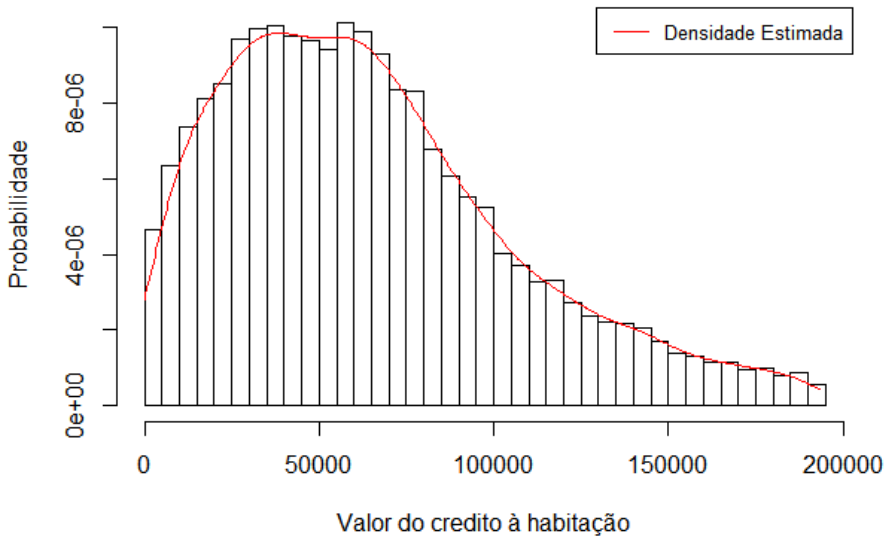
Figura 5.18: Distribuição do valor de crédito vivo (euros/mês).

5.1.4.6 Valor de crédito á habitação

Cerca de 75% dos clientes apresentam créditos á habitação com valores inferiores ou iguais a 100000€. A proporção de clientes que tem créditos á habitação com valores entre 25000 € e 70000 € é idêntica (ver figura 5.19).



(a)

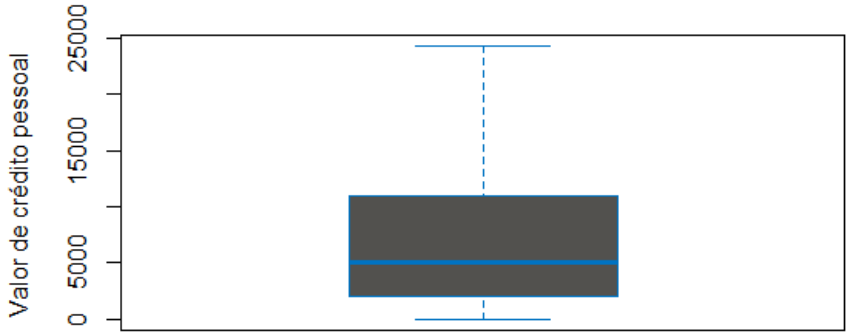


(b)

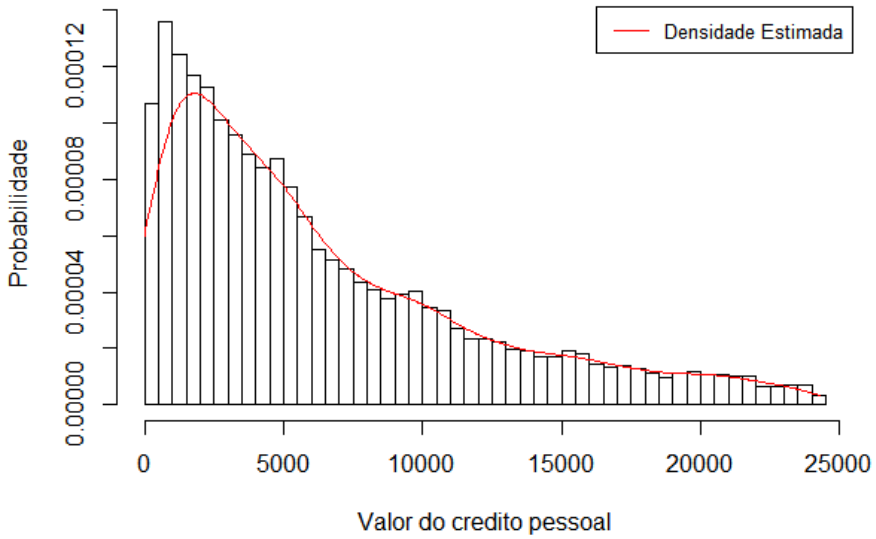
Figura 5.19: Distribuição do valor de crédito à habitação (euros/mês).

5.1.4.7 Valor de crédito pessoal

Verifica-se uma tendência decrescente entre o valor do crédito pessoal e o número de clientes que o detém, ou seja, quanto maior o valor do crédito pessoal, menos clientes o adquirem. Apesar de esta variável apresentar uma mediana que ronda os 5000€, os clientes adquirem mais frequentemente crédito pessoal com valores entre 500€ e 1000€ (ver figura 5.20).



(a)

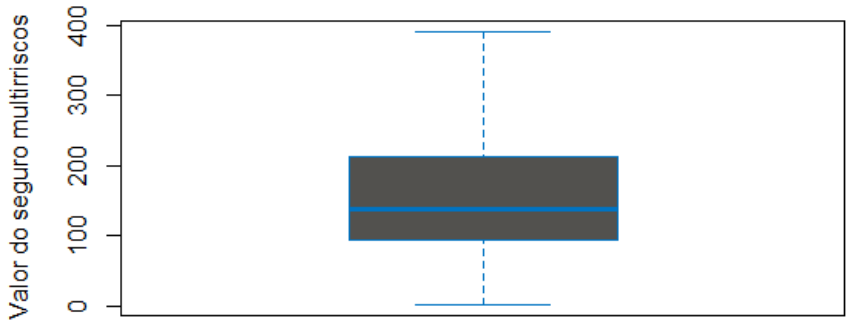


(b)

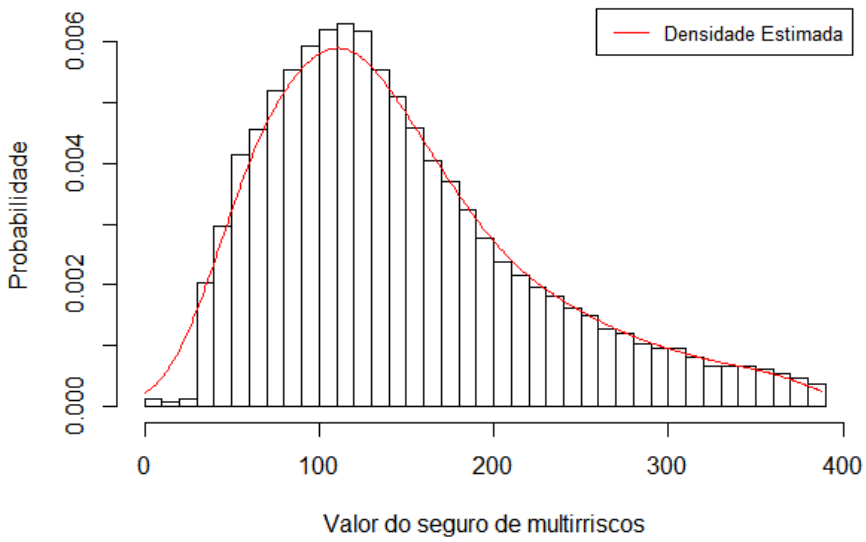
Figura 5.20: Distribuição da densidade do valor de crédito pessoal (euros/mês).

5.1.4.8 Valor do seguro de multirriscos

Analisando a figura 5.21, verifica-se que, é residual a quantidade de clientes que possui seguros multirriscos com valores inferiores a 30€. Pelo *boxplot* desta variável conclui-se que, metade dos clientes deste *dataset* tem seguros multirriscos que variam entre 100 € e 200€.



(a)

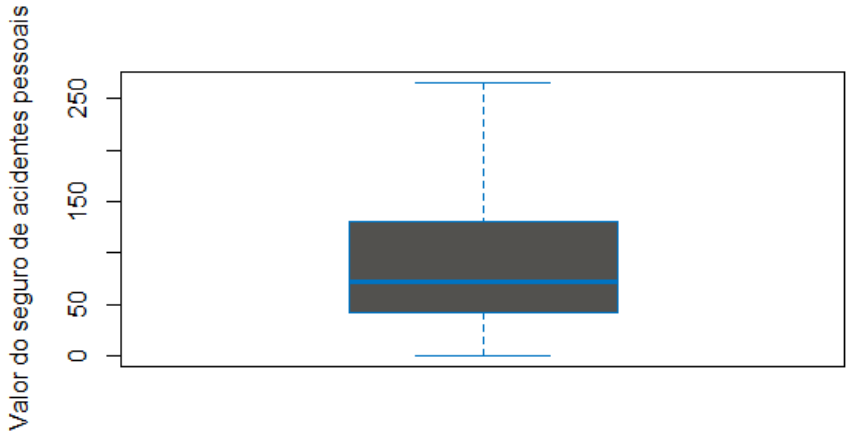


(b)

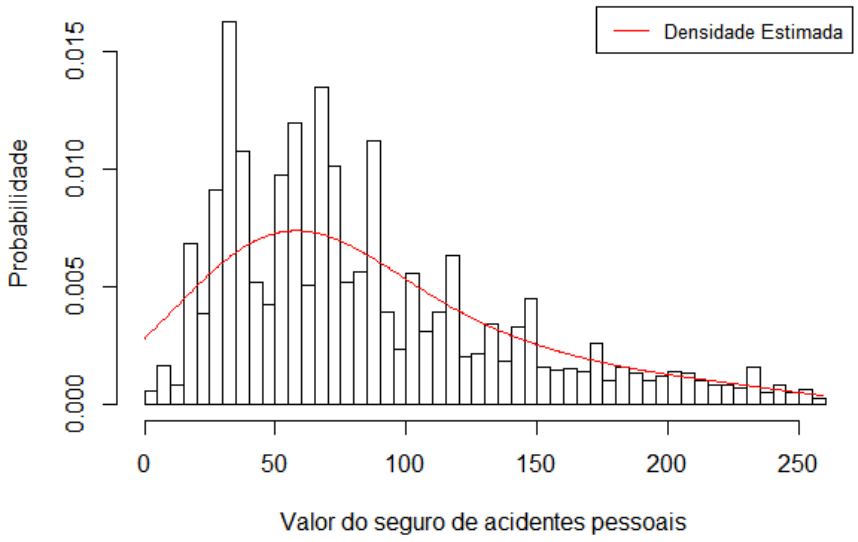
Figura 5.21: Distribuição do valor do seguro multirriscos (euros/mês).

5.1.4.9 Seguro de acidentes pessoais

Através da análise do *boxplot* da figura 5.22, verifica-se que o valor do seguro de acidentes pessoais para metade dos clientes que possuem este produto, é inferior (a cerca de) 80 €. O gráfico da função densidade de probabilidade desta variável apresenta oscilações.



(a)

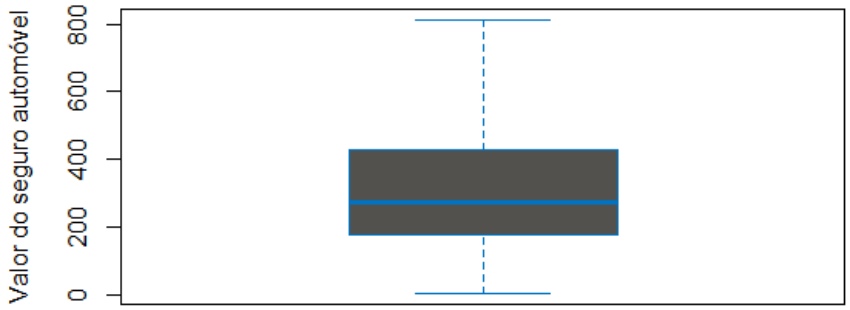


(b)

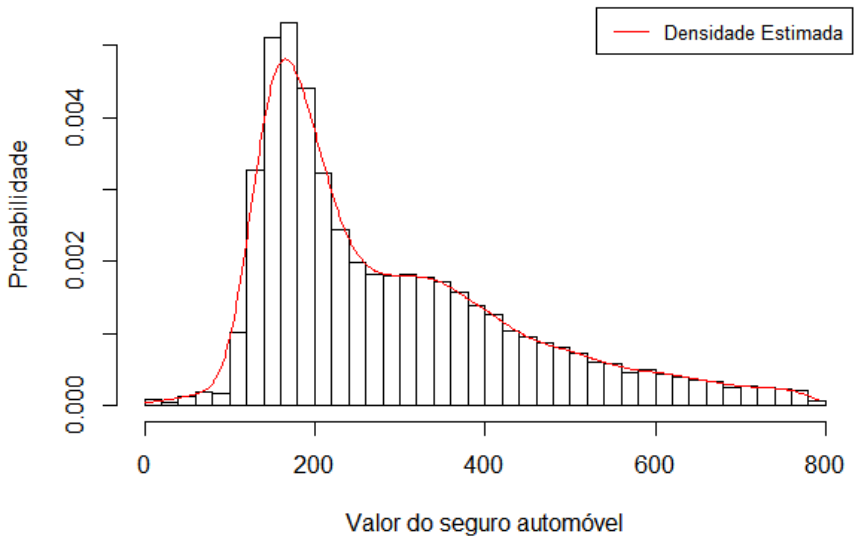
Figura 5.22: Distribuição do valor do seguro de acidentes pessoais (euros/mês).

5.1.4.10 Seguro automóvel

No seguro automóvel e, tal como se verificou no seguro de multirriscos, os clientes que possuem seguros de baixo valor, neste caso, inferiores a 100€ são residuais. Os valores de seguro automóvel mais frequentes variam entre 140€ e 200€. O valor máximo do seguro automóvel registado neste *dataset* foi 800€ (ver figura 5.23).



(a)

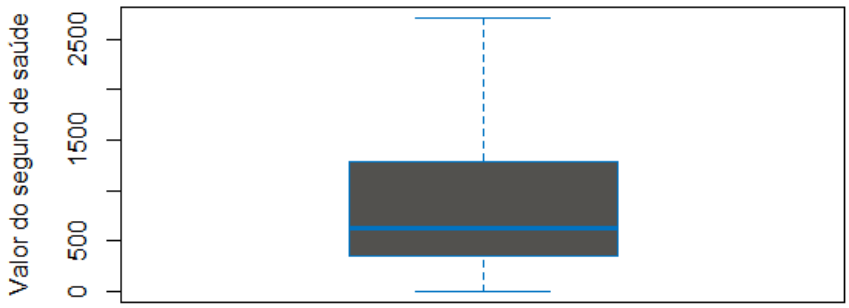


(b)

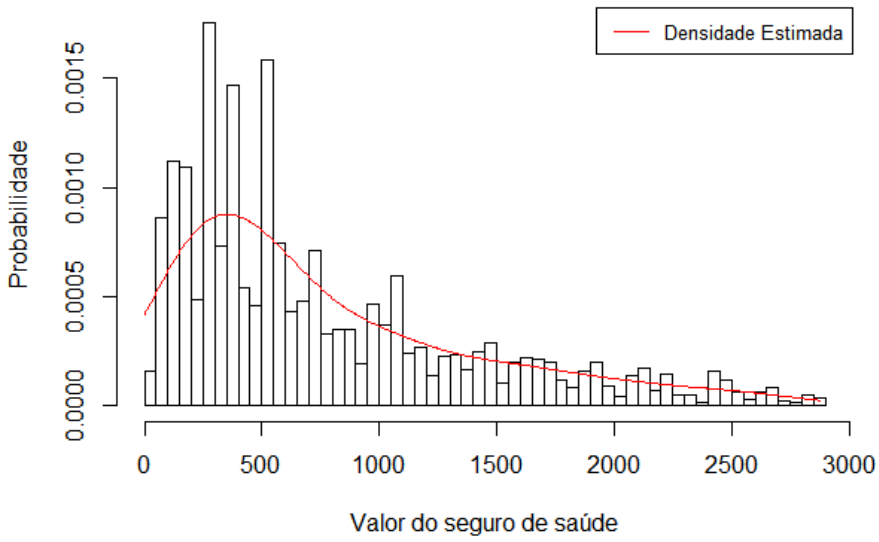
Figura 5.23: Distribuição do valor do seguro automóvel (euros/mês).

5.1.4.11 Seguro de saúde

Esta variável apresenta uma distribuição assimétrica positiva. Da análise do *boxplot* da figura 5.24 conclui-se que metade dos clientes tem seguros de saúde com valores que oscilam entre 500€ e 1500€. O valor máximo registado foi 2500€. Salienta-se que, em regra geral, quanto maior o valor do seguro, menos clientes o possuem.



(a)

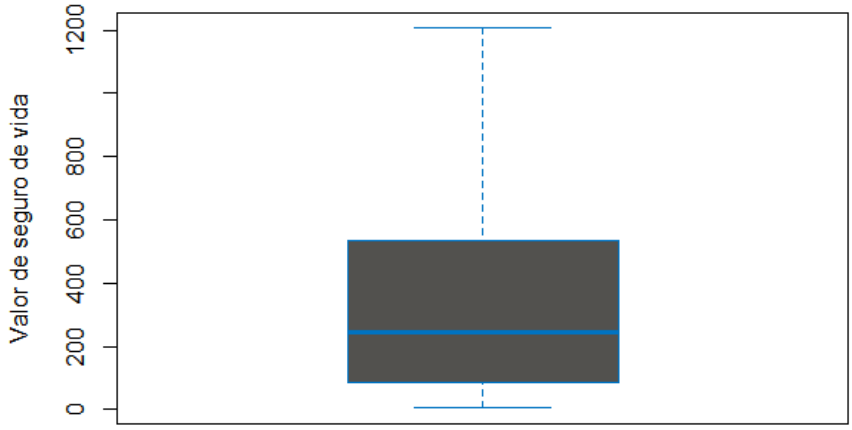


(b)

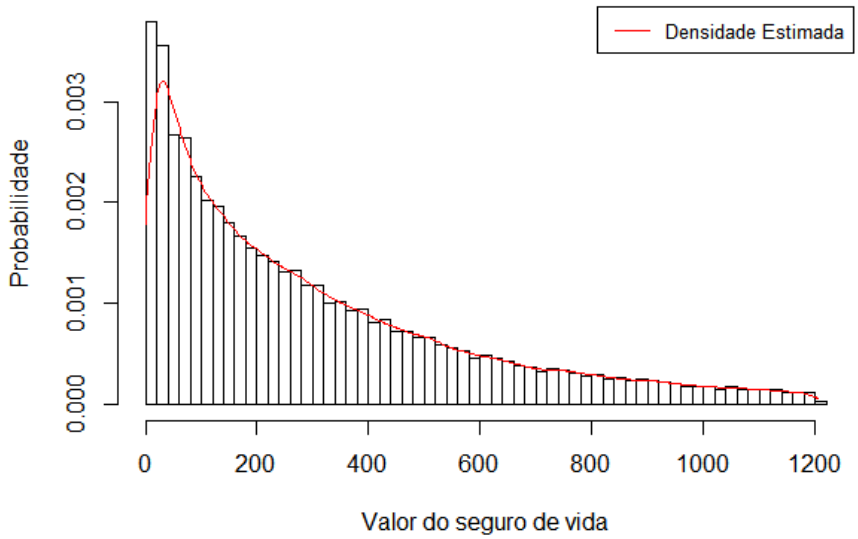
Figura 5.24: Distribuição valor do seguro de saúde (euros/mês).

5.1.4.12 Valor do seguro de vida

Da análise da figura 5.25 conclui-se que esta variável apresenta uma distribuição exponencial e os seguros de vida com valores inferiores a 40 euros são aqueles que os clientes adquirem com mais frequência.



(a)

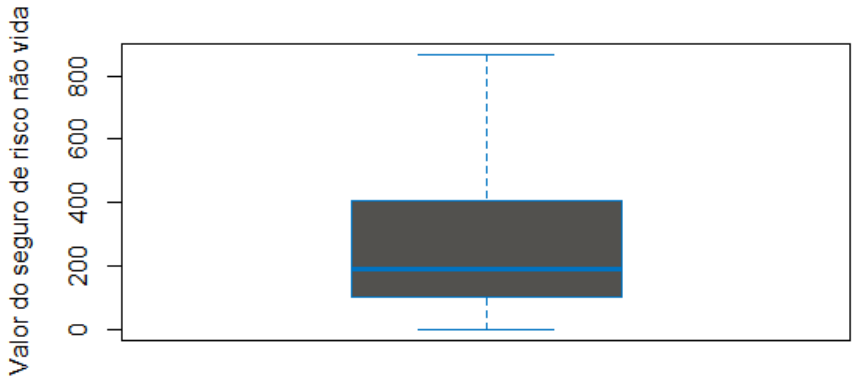


(b)

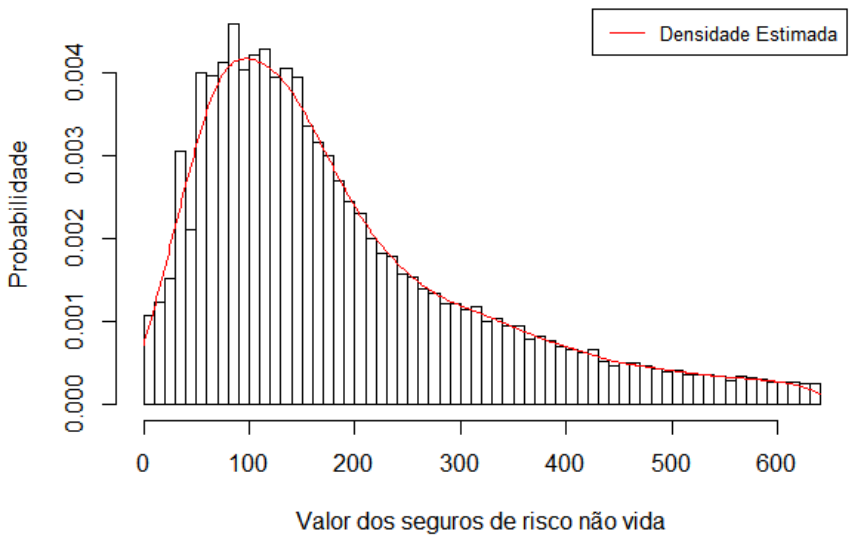
Figura 5.25: Distribuição do valor de seguro de vida (euros/mês).

5.1.4.13 Seguro de risco não vida

O valor máximo verificado para o seguro de risco não vida foi 800€. Verifica-se, também que, 75% dos clientes apresentam seguro de risco não vida com valores inferiores a (cerca de) 400€. Para este tipo de seguro, os valores mais frequentes situam-se entre 70€ e 150€ (ver figura 5.26).



(a)



(b)

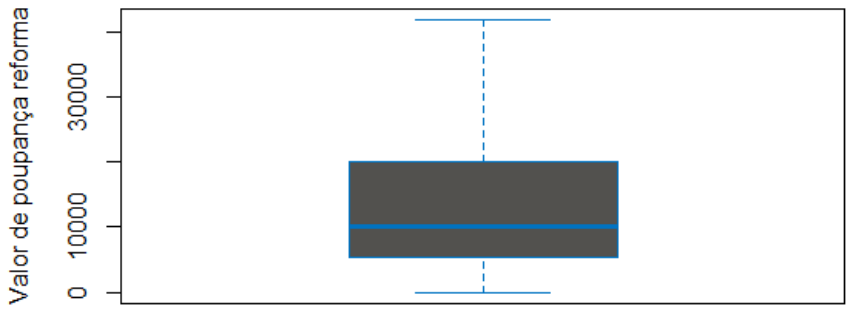
Figura 5.26: Distribuição do valor seguro de risco não vida (euros/mês).

5.1.4.14 Valor da poupança à habitação

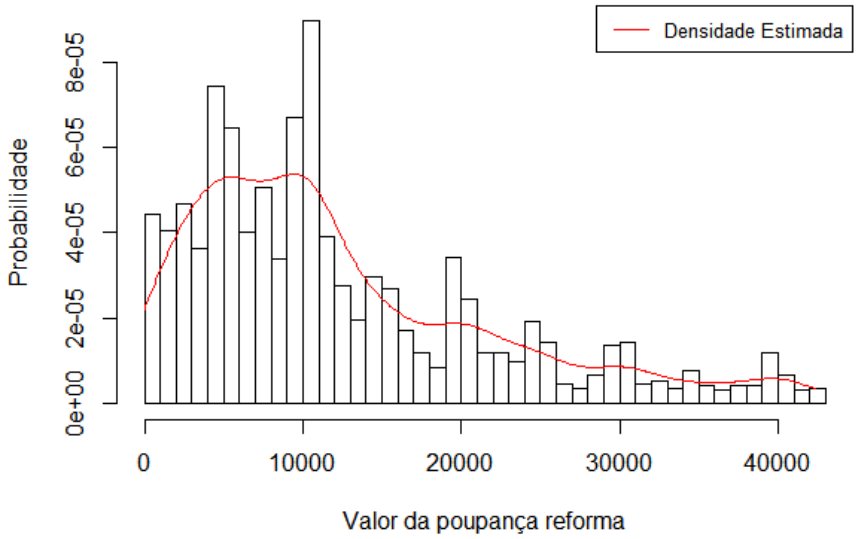
Verifica-se que, para esta variável, existem apenas dois valores possíveis: ou assume valores próximos dos 0€, ou assume valores próximos dos 800€.

5.1.4.15 Valor da poupança reforma

Através da análise do *boxplot* da figura 5.27 conclui-se que, metade dos clientes possuem poupanças reforma com valores que variam entre 5000€ e 17000€. No gráfico de distribuição são perceptíveis algumas oscilações, que provavelmente se devem ao facto de neste *datasets* contarem indivíduos de diferentes classes sociais. Verifica-se um decréscimo na frequência das poupanças reforma com valores superiores a 11000€.



(a)

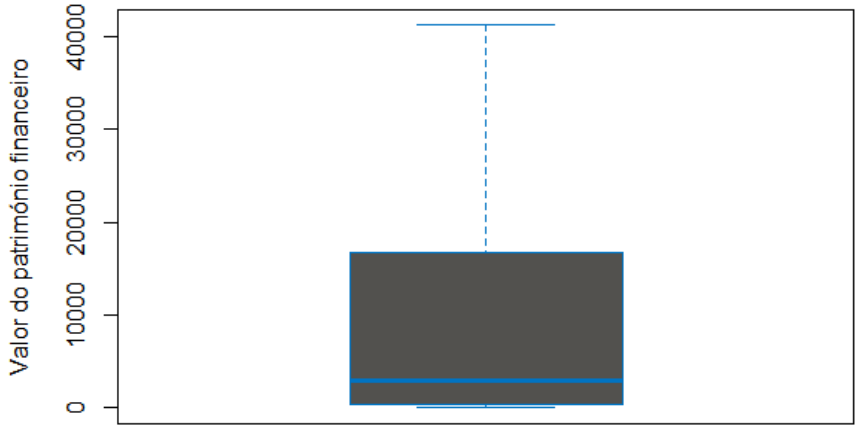


(b)

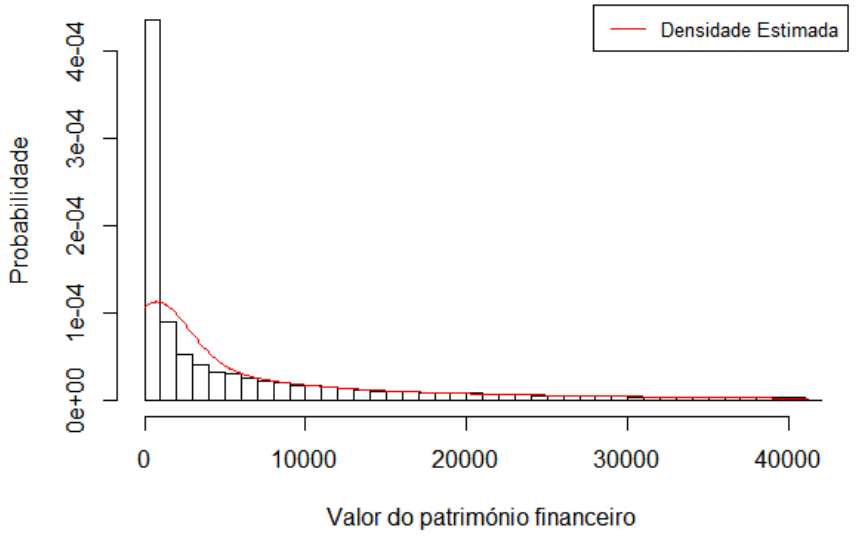
Figura 5.27: Distribuição do valor da poupança reforma (euros/mês)

5.1.4.16 Valor do património financeiro

Analisando o *boxplot* da figura 5.28 verifica-se que, os valores do primeiro quartil e da mediana se encontram relativamente próximos. Esta informação é apoiada pelo gráfico de distribuição, uma vez que, apresenta um comportamento exponencial com decaimento acentuado.



(a)



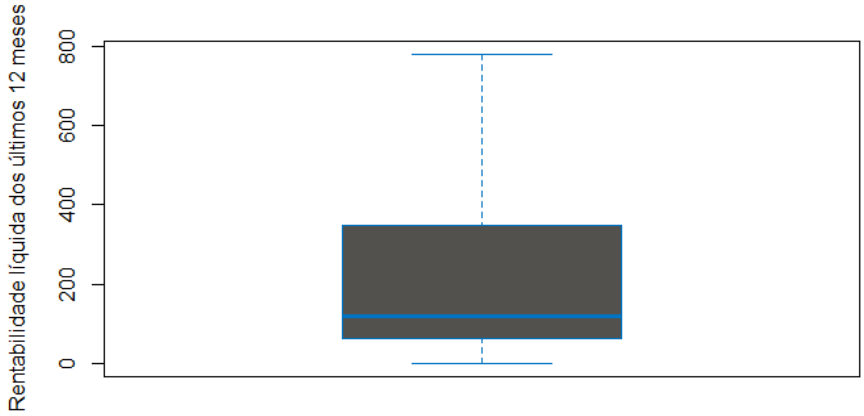
(b)

Figura 5.28: Distribuição do valor do património financeiro (euros).

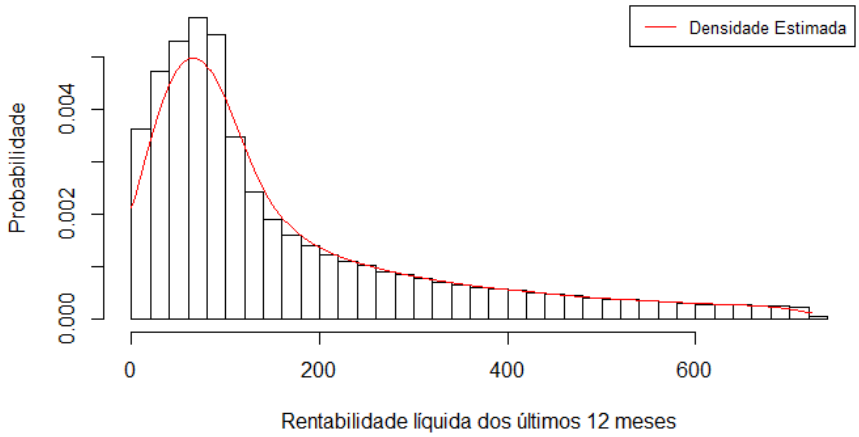
5.1.4.17 Rentabilidade do líquida dos últimos 12 meses

Nas figuras 5.29 e 5.30 encontram-se, representadas, respectivamente, as distribuições dos valores positivos e negativos da Rentabilidade líquida dos últimos 12 meses. Relativamente aos valores positivos desta variável verifica-se que o primeiro e segundo quartil encontram-se relativamente próximos. Neste caso a distribuição é assimétrica positiva. No caso dos valores negativos, verifica-se que, uma maior proximidade entre o segundo e terceiro quartil. Neste caso, a distribuição é assimétrica negativa.

Valores positivos



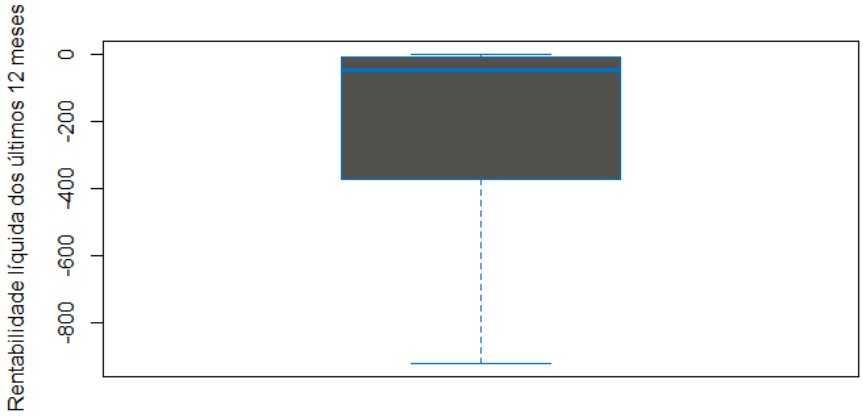
(a)



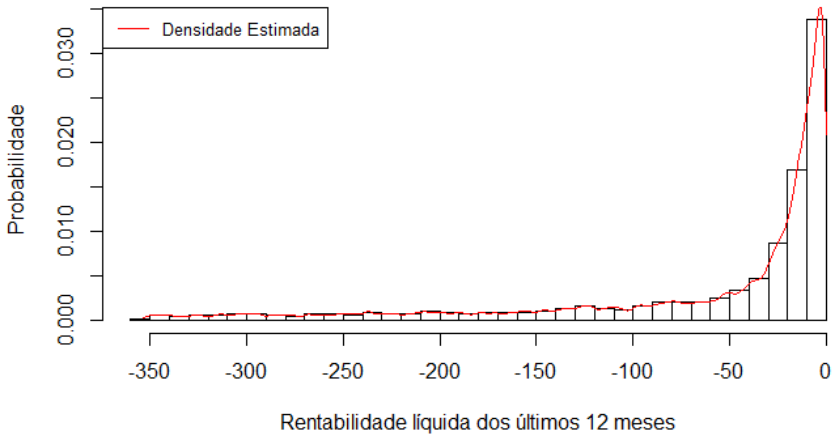
(b)

Figura 5.29: Distribuição do valor da rentabilidade líquida dos últimos 12 meses (anos): valores positivos.

Valores negativos



(a)



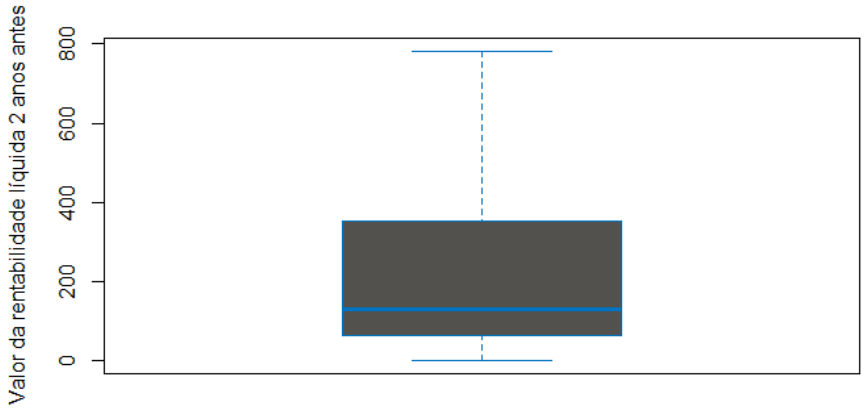
(b)

Figura 5.30: Distribuição do valor da rentabilidade líquida dos últimos 12 meses (anos): valores negativos.

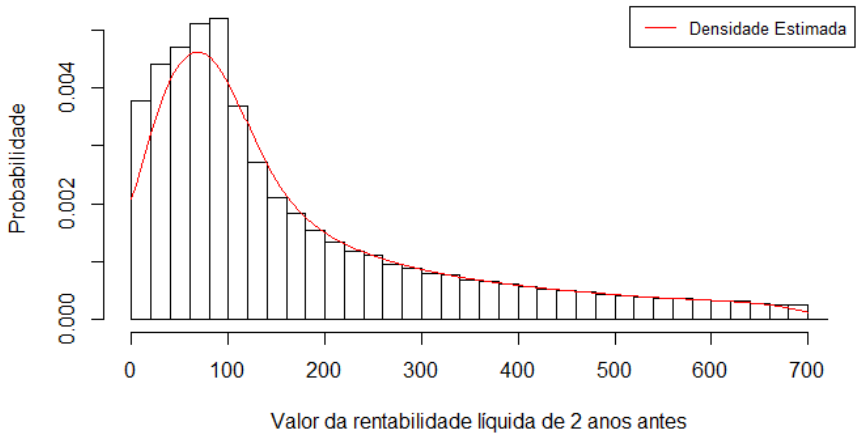
5.1.4.18 Rentabilidade líquida 2 anos antes

A distribuição quer dos valores positivos (figura 5.31), quer dos valores negativos (figura 5.32) da rentabilidade líquida dos últimos 2 anos é idêntica à distribuição da rentabilidade líquida dos últimos 12 meses. Assim, os valores positivos mais frequentes situam-se entre os 60€e os 100€. Os valores negativos mais frequentes situam-se entre os -20 e 0€, ou seja, estes clientes apresentam para o banco um prejuízo na ordem dos 20€.

Valores positivos



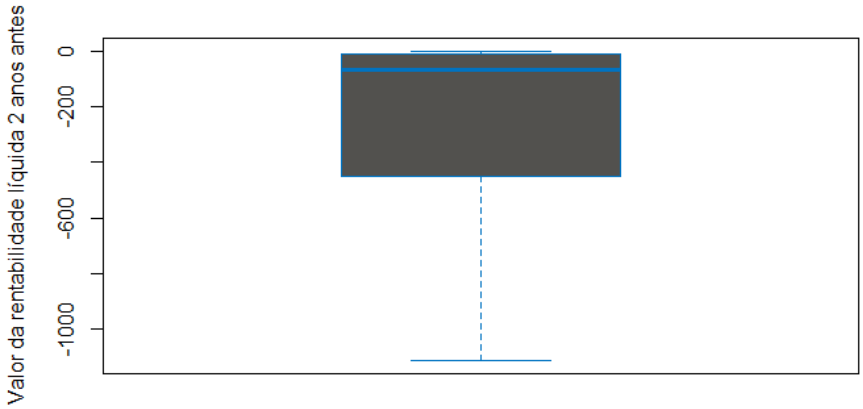
(a)



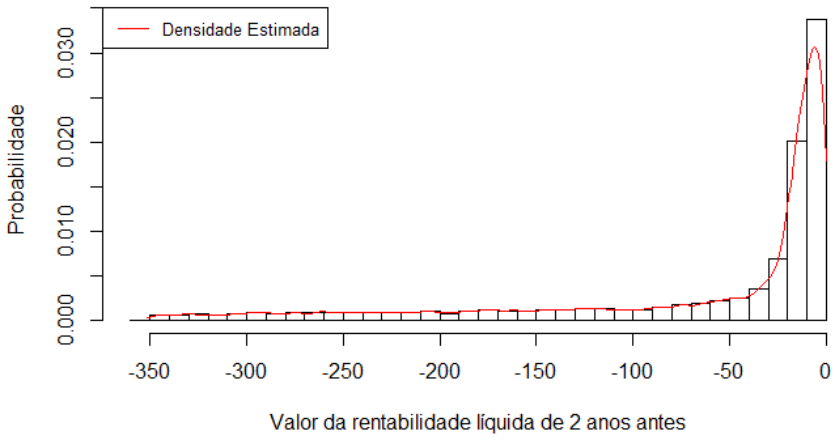
(b)

Figura 5.31: Distribuição do valor da rentabilidade líquida há 2 anos antes (euros): valores positivos.

Valores negativos



(a)

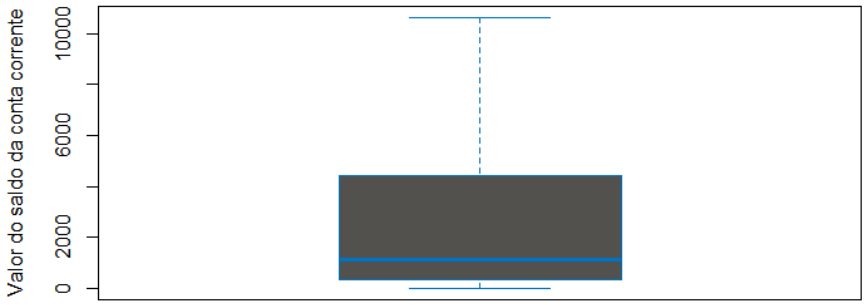


(b)

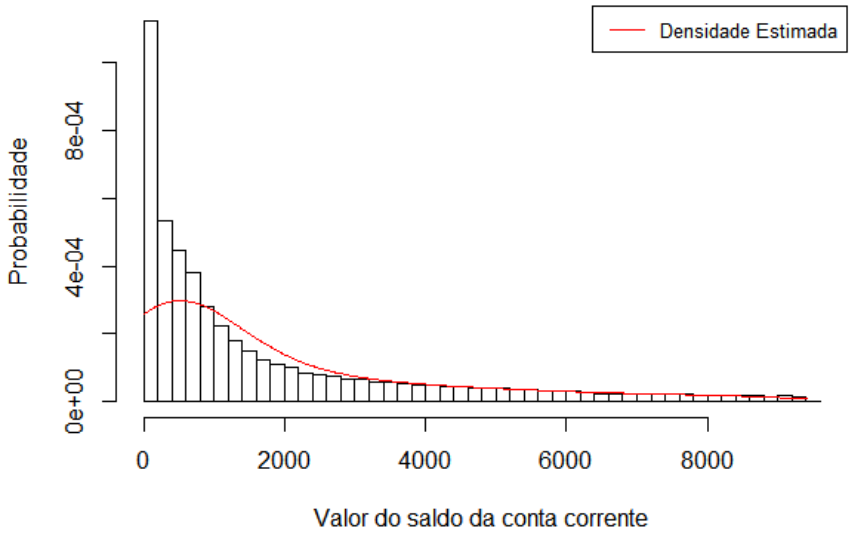
Figura 5.32: Distribuição do valor da rentabilidade líquida há 2 anos antes (euros): valores negativos.

5.1.4.19 Valor do saldo de conta corrente

O valor do saldo da conta corrente (figura 5.33) apresenta uma assimetria positiva, uma vez que a distância interquartis entre o primeiro e o segundo quartil é inferior à distância interquartis existente entre o segundo e o terceiro quartil.



(a)



(b)

Figura 5.33: Distribuição do valor do saldo da conta corrente (euros/mês)

5.1.4.20 Número de cartões

Na análise do número de cartões de crédito que o cliente possui foram tidas em contas duas categorias de cartões: os *Visa* e *Mastercard* e os *American Express*. Verifica-se pouco interesse dos clientes na aquisição deste produto, uma vez que 65.8% dos clientes não possui nenhum cartão de crédito *Visa* e Mastercard. No caso dos cartões *American Express* são 82.6% (ver tabela 5.17).

Tabela 5.17: Distribuição dos cartões *Visa* e *Mastercard* dos cartões *American Express*

Número de clientes	**% de cartões *Visa* e *Mastercard***	**% de cartões *American Express***
0	65.8	82.6
1	24.2	12.7
2	8.8	4.3
≥ 3	1.1	0.5

5.1.5 Variáveis de reclamação

5.1.5.1 *Flag* Reclamação

No mês analisado apenas 192 clientes fizeram pelo menos 1 reclamação, o que corresponde, aproximadamente, a 0.1% dos clientes analisados neste estudo.

5.1.5.2 Dias necessários à resolução da reclamação

A maioria das reclamações (cerca de 89%) foi resolvida no próprio dia e não houve reclamações que demorassem mais de 22 dias a ser resolvidas (ver tabela 5.18).

Tabela 5.18: Dias necessários à resolução da reclamação

Número de dias necessários à resolução	**%**
0	88.5
1	4.2
2	0.5
≥ 3	4.7

5.1.5.3 Valor reclamado

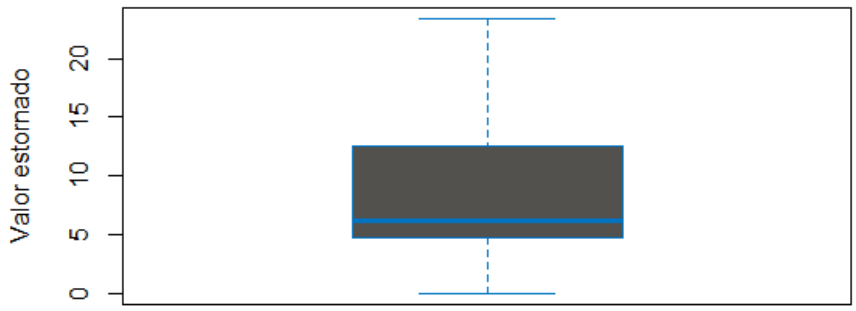
Verifica-se que, 95.9% dos clientes reclamaram valores inferiores a 25€ (tabela 5.19).

Tabela 5.19: Distribuição dos valores reclamados(em euros/mês)

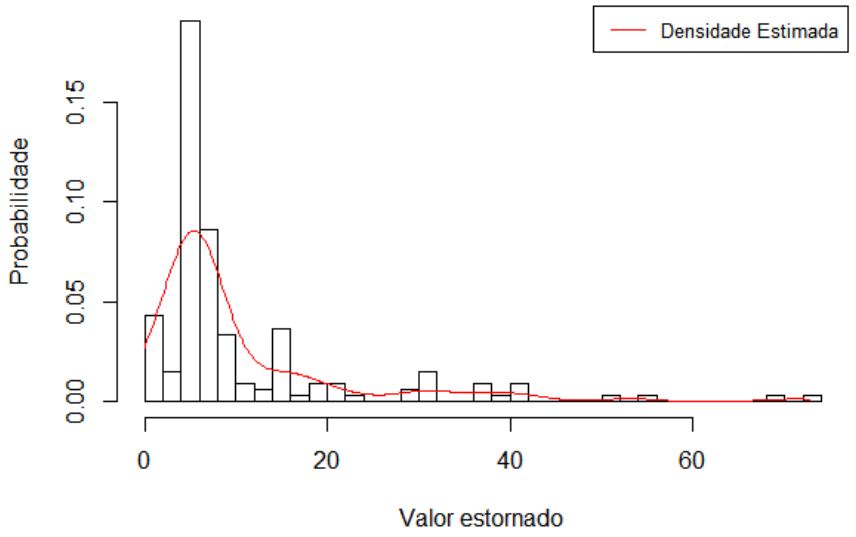
Valor reclamado	**%**
$]0, 25[$	95.9
$[25, 50[$	2.4
≥ 50	1.7

5.1.5.4 Valor estornado

Em caso de reclamação, os valores mais frequentemente estornados variam entre os 4€ e os 6€. Salienta-se que, o valor estornado a 25% dos clientes foi inferior a 5€ (ver figura 5.34) [5].



(a)



(b)

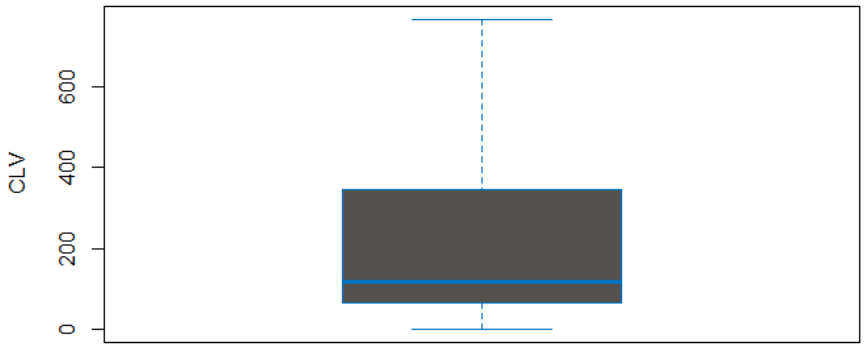
Figura 5.34: Distribuição do valor estornado (euros/mês).

[5] Houve 1 cliente em que o valor estornado foi superior a 100 euros. Esse cliente não se encontra representado no gráfico de distribuição.

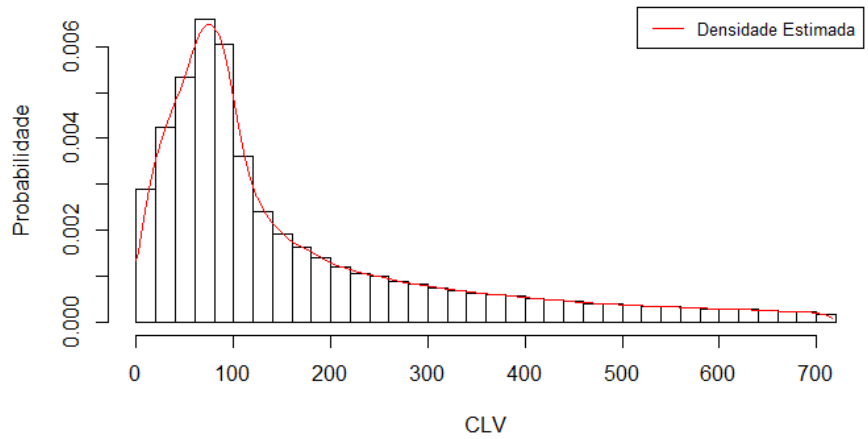
5.1.6 CLV

Nas figuras 5.35 e 5.36 encontram-se as distribuições da variável *target*, o CLV para os seus valores positivos e negativos, respectivamente. Esta variável assume uma distribuição assimétrica positiva para os valores positivos e uma distribuição assimétrica negativa para os valores negativos. Analisando os *boxplots* verifica-se que, existe uma maior distância interquartil entre o mediana e o terceiro quartil, no caso dos valores positivos. Para os valores negativos, a distância entre o primeiro quartil e a mediana é superior às restantes distâncias interquartil.

Valores positivos



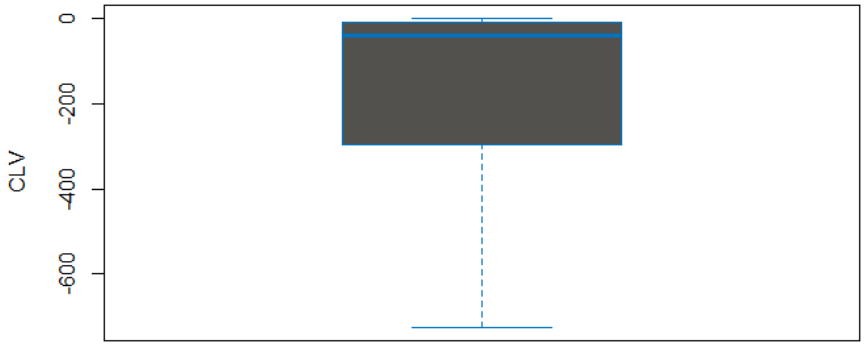
(a)



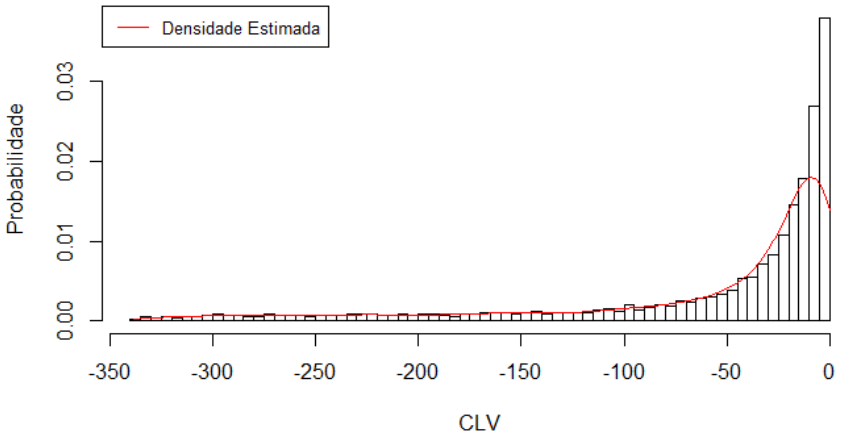
(b)

Figura 5.35: Distribuição do CLV (euros): valores positivos.

Valores negativos



(a)



(b)

Figura 5.36: Distribuição do CLV (euros): valores negativos.

5.2 Modelos

Nesta secção apresentam-se os resultados relativos ao ajustamento dos modelos preditivos do CLV.

5.2.1 CART

Na construção das árvores de decisão foram considerados 2 cenários: 1) Todas as variáveis que constam na tabela 3.1 foram utilizadas como *inputs*; 2) Apenas as variáveis que apresentam importâncias maiores foram utilizadas como *input*.

5.2.1.1 CART com todas as variáveis

Considerando o cenário em que todas as variáveis foram utilizadas como *input* do modelo pode-se afirmar que a rentabilidade dos últimos 12 meses, o valor do crédito vivo, o valor de outros recursos e o montante em produtos (*cross selling*) no ano anterior são as variáveis que aparecem nos nós da árvore, o que significa que, estas são as variáveis que permitem que haja um maior ganho de informação nas partições efectuadas (ver figura 5.37). Analisando as folhas da árvores, conclui-se que, 29% tem um valor de CLV médio de 81€. Esta é folha que contém mais clientes. Nas folhas que correspondem a valores de CLV mais altos, existem poucos clientes.
Considerando todas as variáveis obteve-se uma CART com 9 folhas sendo a rentabilidade dos últimos 12 meses a variável "escolhida" pelo modelo para fazer as primeiras divisões.
Após a construção da CART procedeu-se à sua parametrização. Os parâmetros conside-

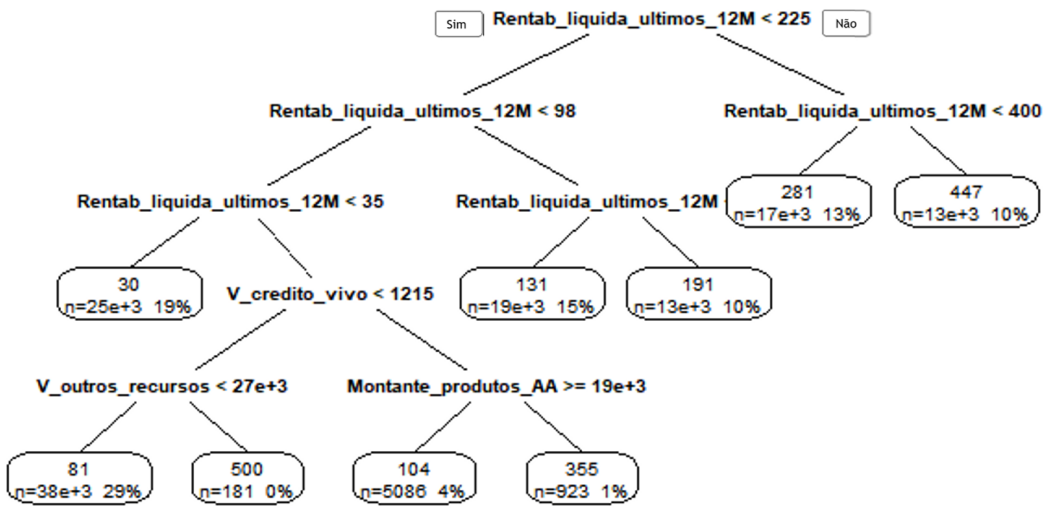


Figura 5.37: Representação do modelo CART gerado utilizando todas as variáveis como *input* (n - número de clientes contidos na folha).

rados na CART apresentada na figura 5.37 foram tal como descrito no capítulo 4, secção

4.2, o *minsplit*, o *maxdepth*, o cp que assumiram os valores 19, 14 e 0.007, respectivamente. Tal como explicado no capítulo 4, secção 4.6, a avaliação do erro foi feita recorrendo ao MAE e ao erro percentual. Neste modelo, para o conjunto de teste o valor do MAE foi 62.7.
Na figura 5.38 encontra-se representada a variação do erro de teste e do erro de treino por decil de CLV. Verifica-se uma diminuição no valor do erro médio absoluto, MAE à medida que os valores do CLV se vão aproximando de zero. Para valores de CLV positivos, o erro absoluto médio,MAE apresenta algumas oscilações: para valores de CLV entre os 0 e 100 € verifica-se uma diminuição no valor do erro; para valores de CLV superiores a 100€ verifica-se um aumento progressivo nos valores do erro médio absoluto.

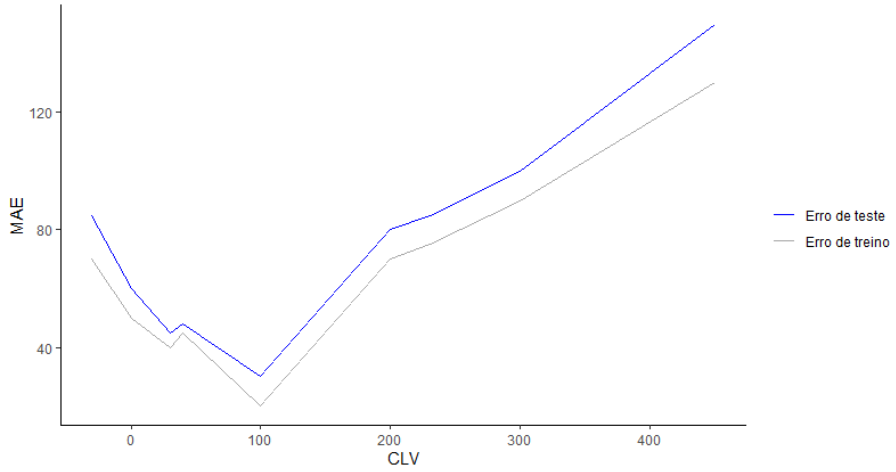


Figura 5.38: Variação do erro absoluto médio (MAE) em função do CLV.

Além do MAE, o erro percentual também foi utilizado para avaliar a precisão do modelo. Na figura 5.39 encontra-se a variação do erro percentual no conjunto de treino e no conjunto de teste por decil. Verifica-se que o erro percentual é superior a 100% para valores de CLV negativos ou próximos de 0€. Após analisar pormenorizadamente estes clientes constatou-se que, em alguns casos o CLV apresentava valores baixos ao longo do tempo, enquanto que em outros casos houve uma queda abrupta no CLV do cliente, que se pode dever a situações de desemprego ou mudança de banco, por exemplo. Esta multiplicidade de cenários e a existência de poucas amostras para cada um dos casos pode explicar a má performance do modelo. Para valores de CLV superiores (a cerca de) 50€ o valor do erro percentual mantém-se constante e apresenta um valor de 5% aproximadamente. No conjunto de teste o erro percentual médio obtido foi 17.25%.
Quer o erro absoluto, quer o erro percentual no conjunto de treino e no conjunto de teste apresentam valores de erro bastante próximos, apesar de o erro no conjunto de teste em ambos os casos ser ligeiramente superior. O MAE e erro percentual obtidos no conjunto de treino foram 62 e 17%, respetivamente.

Na literatura são referidos alguns estudos em que foi utilizado o modelo CART na previsão do CLV. Apesar de algumas variáveis como a antiguidade do cliente ou variáveis relativas ao número de transações terem sido usadas quer no trabalho desenvolvido por Sabben (2018), quer nesta dissertação, houve outras, como por exemplo a propensão do cliente ao *churn*, que Sabben (2018) assumiu como sendo uma variável binária, ou a transacionalidade para países internacionais, que não foram consideradas nesta dissertação. Salienta-se que Sabben não utilizou variáveis demográficas como *input* no seu modelo contrariamente aquilo que foi feito nesta dissertação. A métrica utilizada por Sabben para avaliar a performance do seu modelo foi o erro percentual. Obteve um erro de 10%. Uma vez que a CART construída nesta dissertação apresenta um erro percentual de 17.25%, isto sugere que, a inclusão de variáveis relativas á propensão do cliente ao *churn* poderia ter sido uma escolha assertiva [49].
Rathi (2011) também utilizou o modelo CART para prever o CLV[44]. As variáreis que este autor utilizou como *input* foram a recência, as margens de contribuição de anos anteriores e o número de transações. Nesta dissertação também se optou por usar variáveis referentes aos anos anteriores, nomeadamente a rentabilidade e o montante de produtos. O número de transações é a única variável que foi usada como *input* quer nesta dissertação, quer no trabalho desenvolvido por Rathi. Para avaliar a performance do seu modelo Rathi usou o MAE cujo o erro obtido foi 2343.82. Uma vez que o MAE obtido pelo modelo apresentado nesta dissertação foi 62.7 pode-se concluir que a utilização de um conjunto de variáveis amplo que reflictam o comportamento do cliente melhora a performance do modelo.

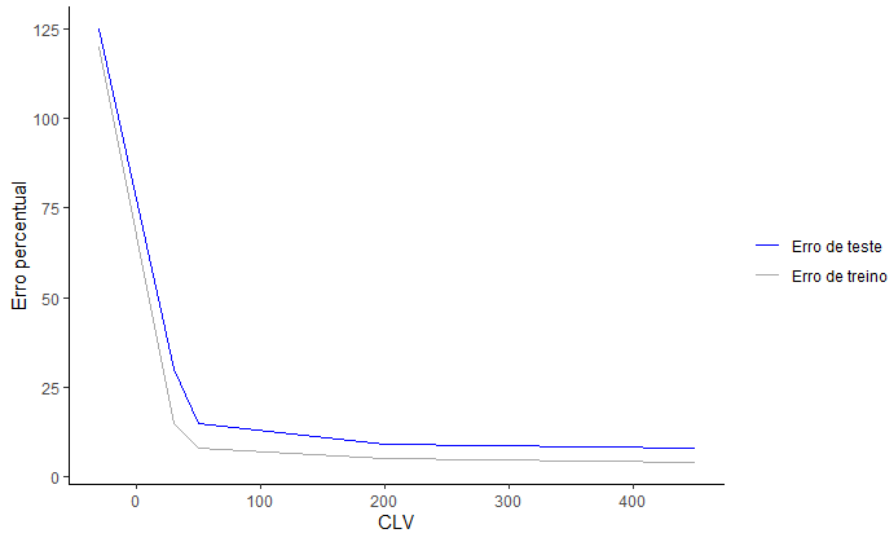


Figura 5.39: Variação do erro percentual em função do CLV.

5.2.1.2 CART com as variáveis mais importantes

De seguida procedeu-se a seleção das variáveis mais importantes e avaliação do modelo nesse cenário. Tal como descrito no capítulo 4, secção 4.2 a métrica utilizada para fazer a seleção foi o ganho de informação (*gain.ratio*). Na figura 5.40 encontram as variáveis que apresentam um ganho de informação superior a 3%. Estas variáveis foram consideradas as mais importantes na previsão e por isso, usadas como *input* no modelo. A variável rentabilidade líquida dos últimos 12 meses destaca-se em relação ás restantes por apresentar um maior ganho de informação.

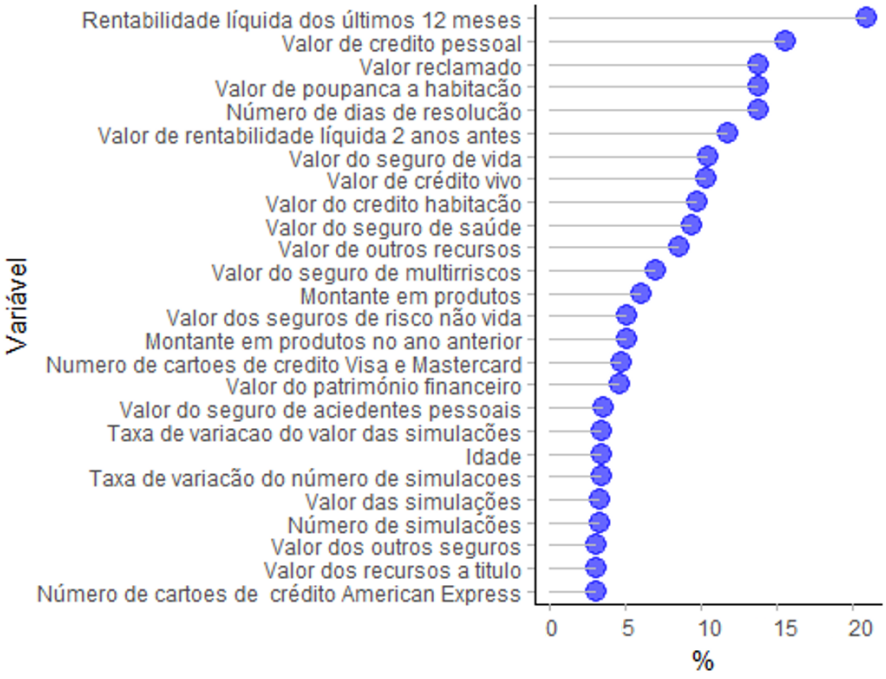


Figura 5.40: Ganho de informação associado a cada variável.

De seguida procedeu-de à construção da CART, à semelhança daquilo que foi feito anteriormente, mas utilizando como *inputs* apenas as variáveis apresentadas na figura 5.40. Na figura 5.41 encontra-se a CART obtida. Comparando esta CART com a CART obtida na figura 5.37 verifica-se que as variáveis utilizadas nos nós são diferentes, enquanto que na CART que utilizou todas as variáveis a rentabilidade líquida dos últimos 12 meses era a variável "escolhida"para efetuar as primeiras partições, no caso da CART gerada utilizando as variáveis mais importantes a variável "escolhida" é a rentabilidade de 2 anos antes. Na CART da figura 5.37 as variáveis valor de outros recursos e montante de produtos do ano anterior são nós, enquanto que na CART da figura 5.41 são utilizadas as variáveis montante de produtos e valor do património financeiro como nós.
Na parametrização da CART foram considerados os parâmetros *minsplit*, *maxdepth* e cp, cujos valores ótimos encontrados foram 19, 14 e 0.005, respectivamente. Para avaliar a performance do modelo foi calculado o MAE e o erro percentual do conjunto de teste e do conjunto de treino por decil. Para o conjunto de teste, os valores do MAE e erro percentual

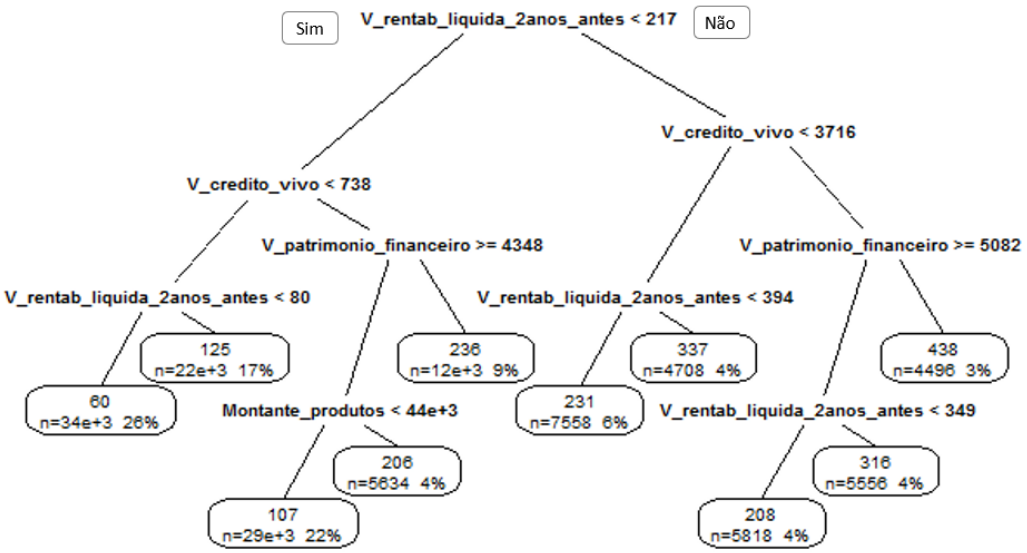


Figura 5.41: Modelo CART gerado utilizando como *input* as variáveis que apresentam um ganho de informação superior a 3% (n - número de clientes contidos na folha).

obtidos foram 87.44 e 33.60%, respetivamente. No conjunto de treino, obteve-se um MAE de 87.22 e um erro percentual de 33.75%. Assim, pode-se concluir que a performance da CART que utilizou todas as variáveis como *input* é superior à da CART que utilizou apenas as variáveis que apresentavam um ganho de informação superior a 3% como *input*. Nas figura 5.42 e 5.43 está representada a variação do MAE por decil e a variação do erro percentual por decil, respetivamente. Em ambas as figuras verifica-se que apresentam um comportamento similar ao das suas homologas apresentadas no cenário em que foram consideradas todas as variáveis como *input*, apesar de neste caso, os valores dos erros em ambas as figuras serem superiores. A análise do erro por decil de CLV também permite concluir que, a "faixa" de valores de CLV que apresenta valores de erro médio absoluto mais baixos é entre os 100 e 120€. Para valores negativos, o erro médio absoluto, MAE, ronda os 130€, esse valor diminui progressivamente até se atingir um valor de CLV de 100€. Para valores de CLV superiores a 120€ verifica-se, progressivamente um aumento no valor do erro médio absoluto (MAE). O erro percentual permanece aproximadamente constante para valores de CLV superiores a (cerca de) 50€.

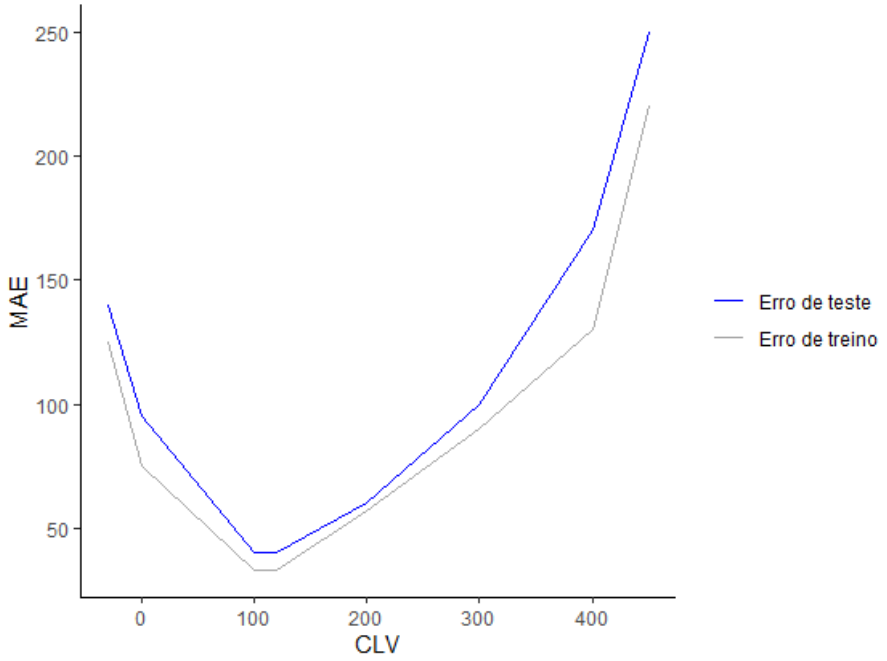


Figura 5.42: MAE por decil

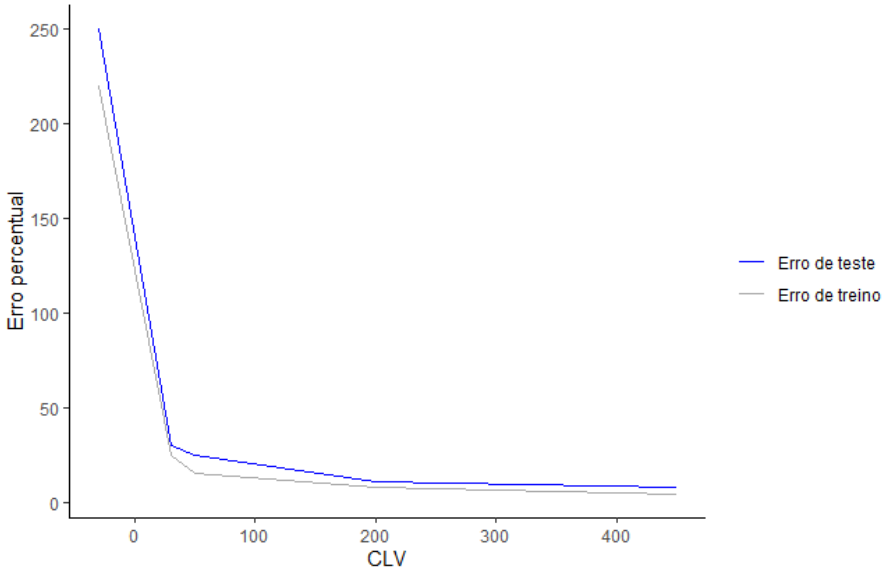


Figura 5.43: Variação do erro absoluto médio (MAE) em função do CLV.

5.2.2 Random Forest

O modelo RF foi implementado com o intuito de melhorar os resultados obtidos pelo modelo CART, uma vez que, este método apresenta uma maior robustez. Na implementação da *Random Forest*, e, à semelhança daquilo que foi feito na CART, foram considerados dois cenários: 1) Todas as variáveis do *dataset*, descritas na tabela 3.1, foram consideradas como variáveis de *inputs*; 2) Apenas as variáveis mais importantes, ou seja, as que apresentam uma maior explicabilidade do *output*, foram consideradas como *inputs*.

5.2.2.1 RF com todas as variáveis

Na construção da RF o número de árvores foi um dos parâmetros considerados, tal como descrito no capítulo 4, secção 4.3. De modo a encontrar o valor ótimo para esse parâmetro fez-se o *plot* do erro quadrático médio[6] em função do número de árvores consideradas. Verifica-se uma descida exponencial do erro quadrático médio até ao momento em que são consideradas cerca de 40 árvores para o treino do algoritmo (ver figura 5.44). Considerando mais de 40 árvores, apenas se verifica uma ligeira descida do erro. Por esse motivo, o número de árvores considerado nesta implementação foi 40, uma vez que se considerou que o ganho em termos de diminuição do erro era pequeno, não justificando o aumento da complexidade do modelo.

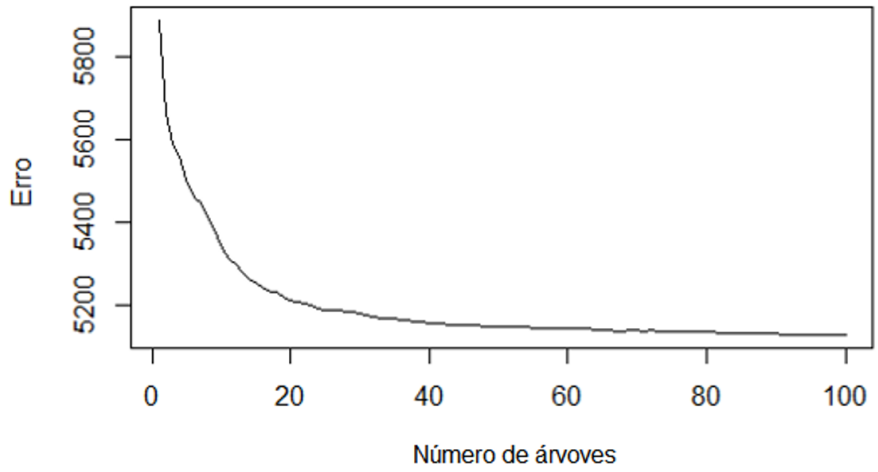


Figura 5.44: Variação do erro quadrático médio em função do número de árvores consideradas

Neste estudo foram utilizados mais 2 parâmetros para otimizar os resultados da *Random Forest*: o *nodesize* e o *maxnodes*. Os valores atribuídos a estes parâmetros foram, respetivamente, 5 e 4.
De seguida procedeu-se á avaliação da performance do modelo. Para esse fim foram utilizadas duas métricas: o MAE e o erro percentual, tal como descrito no capítulo 4, secção 4.6. No conjunto de teste obteve-se um MAE de 47.90 e um erro percentual de 12.68%.

[6] O erro quadrático médio é definido pela diferença entre o valor estimado e o target ao quadrado.

No conjunto de treino o MAE foi 20.33 e o erro percentual foi 5.98%.
Na figura 5.45 encontra-se a variação do MAE no conjunto de treino e no conjunto de teste por decil de CLV. Verifica-se que independentemente do valor do CLV, o erro de teste é sempre mais alto do que o erro de treino.

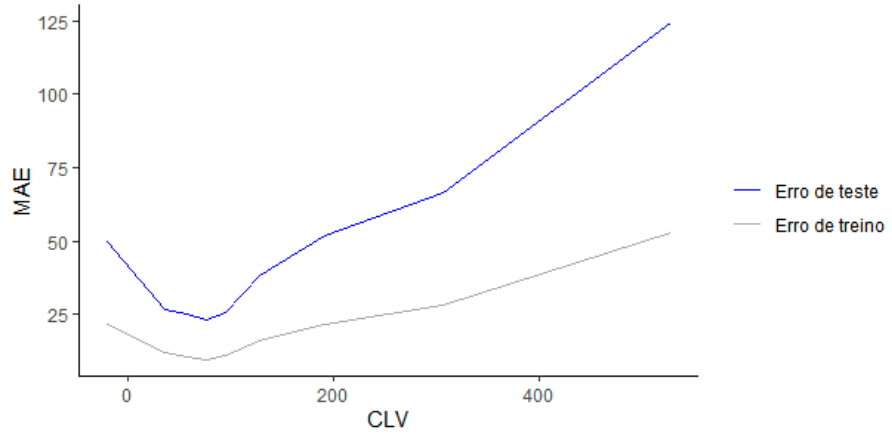


Figura 5.45: Variação do erro absoluto médio(MAE) em função do CLV.

Na figura 5.46 encontra-se a variação do erro percentual no conjunto de treino e no conjunto de teste por decil de CLV. O erro de treino mantém-se sempre mais baixo do que o erro de teste. Quando o CLV apresenta valores inferiores a 0, ou próximos de 0, o erro percentual é mais elevado. Para valores de CLV superiores a (cerca de) 30€ o erro percentual mantém-se aproximadamente constante e com valores próximos de zero.

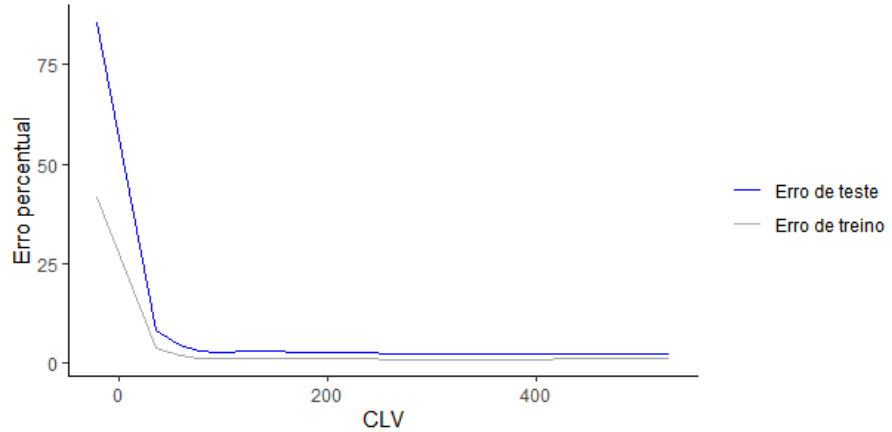


Figura 5.46: Erro percentual por decil.

Comparando estes resultados com aqueles que foram obtidos na literatura, verifica-se que a utilização das RF em vez das CARTs se traduz numa melhoria de performance. Sabbeh (2018) que, no seu estudo, para além das CARTs também usou as RF e, verificou

uma melhoria dos seus resultados após aplicar as RF, tendo o erro percentual diminuído para 3.7% [49].

5.2.2.2 RF com as variáveis mais importantes

De seguida, e à semelhança daquilo que foi feito nas CARTs, procedeu-se à construção das RF tendo por base apenas as variáveis mais importantes. Definiu-se como valor de *threshold* aquele em que se verificava uma maior "quebra de importância", por uma questão de equidade em relação ao procedimento adoptado no modelo CART. Na figura 5.47 são apresentadas as variáveis que foram usadas como *input* no modelo, bem como as respetivas percentagens de IncMSE.

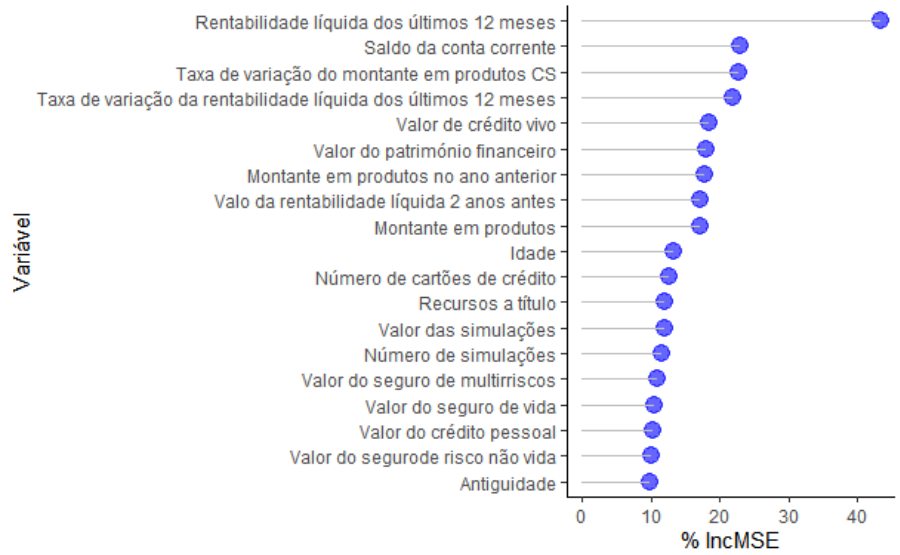


Figura 5.47: %IncMSE de cada variável.

Neste cenário, obteve-se um MAE de 48.21 e um erro percentual de 13.41%, para o conjunto de teste, ou seja, verificou-se uma pior performance nas RF que tiveram apenas as variáveis mais importantes como *input*. No conjunto de treino registou-se um MAE de 20.59 e um erro percentual de 6.28%.
Salienta-se que na construção desta RF utilizaram-se os mesmos parâmetros que foram apresentados, em cima, para a construção da RF que utilizou todas as variáveis como *input*, por se verificar, também neste caso, que correspondiam aos parâmetros ótimos. Nas figuras 5.48 e 5.49 encontram-se representada a variação do erro de treino e do erro de teste por decis de CLV utilizando o MAE e o erro percentual, respectivamente. Quer no MAE, quer no erro percentual a variação do erro em função do CLV apresenta a mesma tendência verificada anteriormente quando foram consideradas todas as variáveis como *inputs*, apesar de, neste caso, os valores do MAE e erro percentual serem ligeiramente superiores.

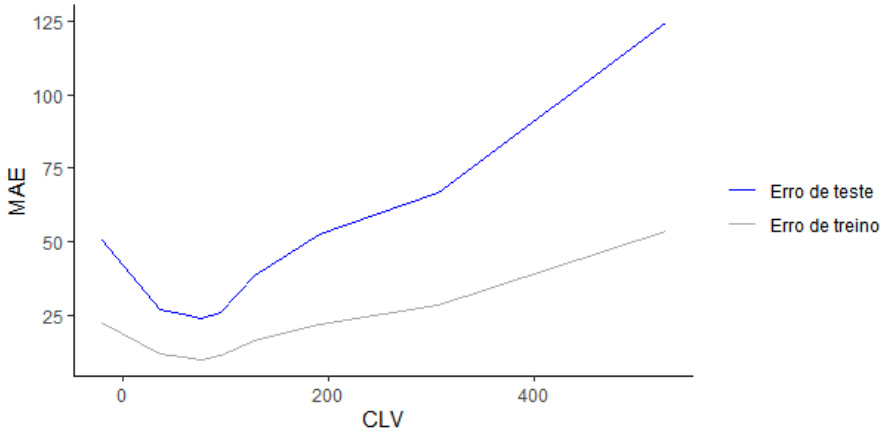


Figura 5.48: MAE por decil.

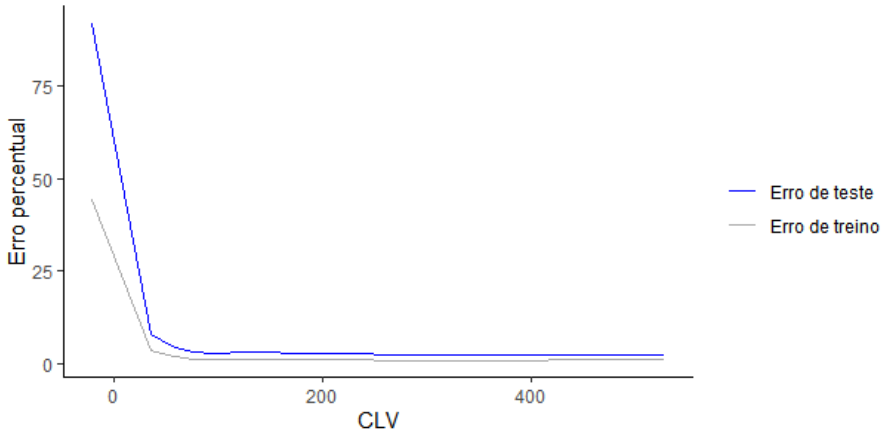


Figura 5.49: Erro percentual por decil.

5.2.3 Cadeias de Markov

No desenvolvimento desta abordagem, em que se utilizaram as cadeias de Markov para a predição do CLV, foi tida em consideração a abordagem de Haenlein (2007), tendo inclusive, sido construída uma base de dados para esta implementação que foi ao encontro das variáveis que Haenlein (2007) utilizou no seu estudo [22]. O autor utilizou um modelo CART para determinar os estados das cadeias de Markov, conseguindo dessa forma obter grupos homogéneos de clientes. No *dataset* utilizado por Haenlein, a idade era a variável que apresentava maior importância. Por esse motivo, o autor decidiu construir o modelo CART tendo em conta intervalos etários, ou seja, separou os clientes por faixa etária e de seguida implementou o modelo para cada faixa etária. No caso do *dataset* utilizado nesta dissertação, a variável que apresenta maior relevância era rentabilidade dos últmos 12 meses e, por isso, decidiu-se fazer a partição dos clientes tendo em conta esta variável. Assim, os clientes foram divididos em 3 grupos:

- **Grupo 1**: Clientes cuja rentabilidade líquida dos últimos 12 meses fosse superior ao terceiro quartil mas não fosse considerada *outlier*.
- **Grupo 2**: Clientes que apresentavam rentabilidade líquida dos últimos12 meses entre o primeiro e o terceiro quartil;
- **Grupo 3**: Clientes cuja rentabilidade líquida dos últimos 12 meses fosse inferior ao primeiro quartil mas não fosse *outlier*;

Segue-se a apresentação das árvores de decisão obtidas para cada um dos intervalos de rentabilidade considerados, bem como a análise feita com o intuito de escolher o número ideal de estados para cada cenário. Salienta-se que o número de estados considerados corresponde ao número de folhas existentes no modelo CART. Este método foi descrito no capítulo 4, secção 4.2
No processo de poda foram testados vários valores para o parâmetro de complexidade (cp) e, de seguida calculados os respetivos erros. A árvore em que se verificava uma melhor relação entre o seu tamanho e capacidade preditiva era a escolhida para definir os estados.

5.2.3.1 Definição dos estados

Na definição dos estados das Cadeias de Markov foi testado o ganho associado à inserção da rentabilidade líquida dos últimos 12 meses como *feature*. A principal vantagem da utilização da rentabilidade líquida dos últimos 12 meses é o facto de esta variável apresentar uma grande importância, e consequentemente melhorar a capacidade preditiva do modelo. No entanto, esta *feature* também acaba por "camuflar" as restantes, não permitindo saber que outras variáveis são importantes e tem relevância na predição do CLV de cada um dos grupos.
Na tabela 5.20 encontram-se os menores valores do erro de teste obtido pelas CART para

cada um dos grupos. Foi considerado o cenário em que a rentabilidade líquida dos últimos 12 meses foi incluída no *set* de *input* e o cenário em que essa variável foi excluída do *set* de *input*. Como as diferenças nos valores do erro não foram significativos e como os quartis da rentabilidade líquida dos últimos 12 meses foram utilizados para fazer a separação dos clientes nos diferentes grupos, optou-se por não incluir esta variável como *input*.

De seguida serão apresentadas os modelos CART considerados na definição dos estados de cada grupo e, as parametrizações efetuadas para cada um dos casos.

Na escolha do modelo CART "ótimo" para definir os estados de cada um dos grupos foram tidos em conta 2 critérios: a eficácia preditiva do modelo e o número de folhas, uma vez que, o número de folhas irá corresponder ao número de estados considerados nas Cadeias de Markov. Convêm referir que, quanto maior for o número de estados considerados, maior será a complexidade associada ao cálculo da matriz de transição.

De modo a obter a árvore que melhor se adequa ao objectivo foram testados vários valores do parâmetro de complexidade (cp). Este parâmetro interfere directamente no tamanho das árvores e, consequentemente, no número de folhas existentes nas árvores, ou seja, aumentando o valor do parâmetro de complexidade ocorre uma diminuição do tamanho da árvore. Caso se diminua o valor do parâmetro de complexidade o tamanho da árvore aumentará.

Os modelos CART em que a rentabilidade do ano anterior foi considerada como *input* encontram-se no apêndice A.

Tabela 5.20: Erro médio absoluto obtido para cada um dos grupos em função da inclusão ou exclusão da rentabilidade do ano anterior como variável de *input*

Grupo	**Erro de teste incluindo a rentabilidade do ano anterior como *input***	**Erro de teste excluindo a rentabilidade do ano anterior *input***
Grupo 1	63.8	79.6
Grupo 2	18.9	26.9
Grupo 3	14.2	14.2

CARTs para valores superiores ao terceiro quartil de rentabilidade dos últimos 12 meses (Grupo 1)

Neste grupo foram gerados 3 modelos CART, que irão ser designados por CART 1, CART 2 e CART 3. Na tabela 5.21 apresenta-se o erro de teste (calculado utilizando o *Mean Absolute Error*), o número de folhas e o parâmetro de complexidade (cp) associado a cada um dos modelos CART. Comparando a complexidade das 3 árvore geradas e os seus respetivos erros verifica-se que, o erro mais baixo é obtido para a árvore com mais folhas (CART 3). Comparando o erro obtido por esta árvore com o erro obtido pela árvore que tem apenas 3 folhas (CART 1), a diferença é pouco significativa, e por isso, optou-se pela CART 1, representada na figura 5.50, para definir os estados da Cadeia de Markov. A CART2 com apenas 2 folhas não é a melhor alternativa porque apresenta um erro superior ao das restantes CARTs.

Tabela 5.21: Erro médio absoluto (erro de teste) em função do número de folhas e parâmetro de complexidade (cp)

CART	Erro de teste	Número de folhas	Parâmetro de complexidade (cp)
CART 1	79.6	3	1×10^{-2}
CART 2	82.4	2	8×10^{-2}
CART 3	78.1	6	7×10^{-3}

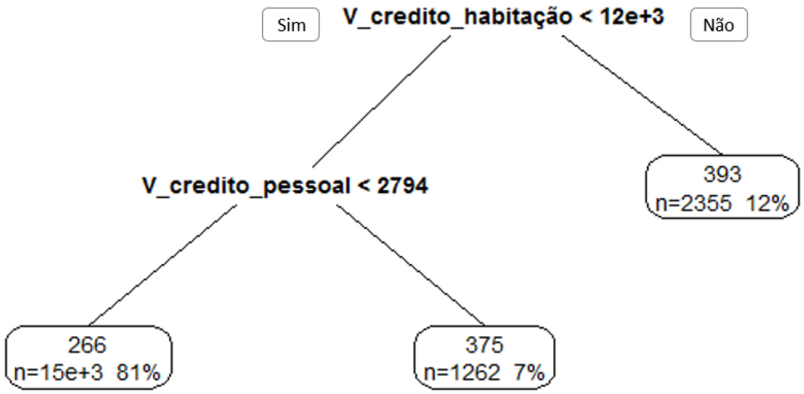


Figura 5.50: Modelo CART 1 (n - Número de clientes alocados à folha).

CARTs para valores de rentabilidade dos últimos 12 meses entre o primeiro e o terceiro quartil (Grupo 2)

Neste grupo foram gerados 2 modelos CART, que irão ser designados por CART 4 e CART 5. Na tabela 5.22 apresenta-se o erro médo absoluto (teste), o número de folhas e o parâmetro de complexidade (cp) associado a cada um dos modelos CART obtidos para o intervalo de rentabilidade do ano anterior entre o 1º e o 3º quartil.

Analisando as duas árvores obtidas para este intervalo de rentabilidade dos últimos 12 meses, a melhor escolha recai sobre o modelo CART 5 (figura 5.51) uma vez que, esta apresenta menos estados e o erro das duas árvores é idêntico.

Tabela 5.22: Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp)

CART	Erro de teste	Número de folhas	Parâmetro de complexidade (cp)
CART 4	26.7	4	1×10^{-2}
CART 5	26.9	3	2.4×10^{-2}

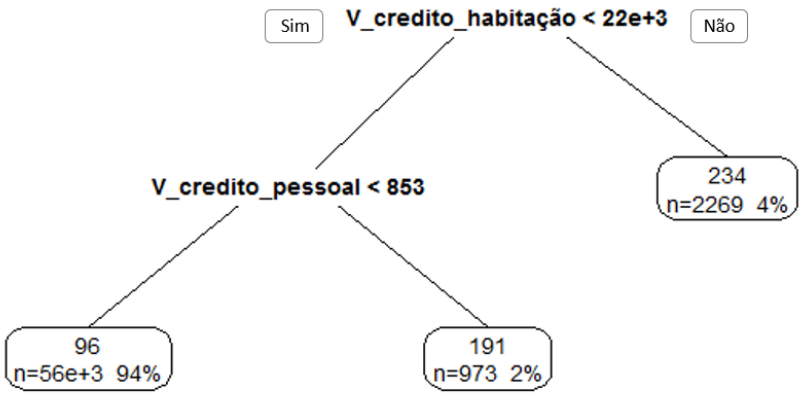


Figura 5.51: Modelo CART 5 (n - Número de clientes alocados à folha).

CARTs para valores de rentabilidade dos últimos 12 meses inferiores ao primeiro quartil (Grupo 3)

Neste grupo foram gerados 4 modelos CART, que irão ser designados por CART 6, CART 7, CART 8 e CART 9. Na tabela 5.23 apresenta-se o erro médio absoluto (teste), o número de folhas e o parâmetro de complexidade (cp) associado a cada um dos modelos CART geradas para o intervalo em que a rentabilidade dos últimos 12 meses é inferior ao 1º

quartil.
O modelo CART 7 apresenta um erro superior ao dos restantes modelos CART. Os valores dos erros obtidos para os modelos CART 6, CART 8 e CART 9 são muito idênticos, por esse motivo, o critério de escolha entre estas três árvores foi o número de folhas. O modelo CART 6 (figura 5.52) é a que apresenta menor número de folhas e por esse motivo, foi o escolhido para definir os estados para este grupo.

Tabela 5.23: Erro médio absoluto (erro de teste) em função do número de folhas e parâmetro de complexidade (cp)

CART	Erro de teste	Número de folhas	Parâmetro de complexidade (cp)
CART 6	14.2	3	1.0×10^{-2}
CART 7	15.2	2	3.3×10^{-2}
CART 8	14.1	5	7.0×10^{-3}
CART 9	14.1	4	7.7×10^{-3}

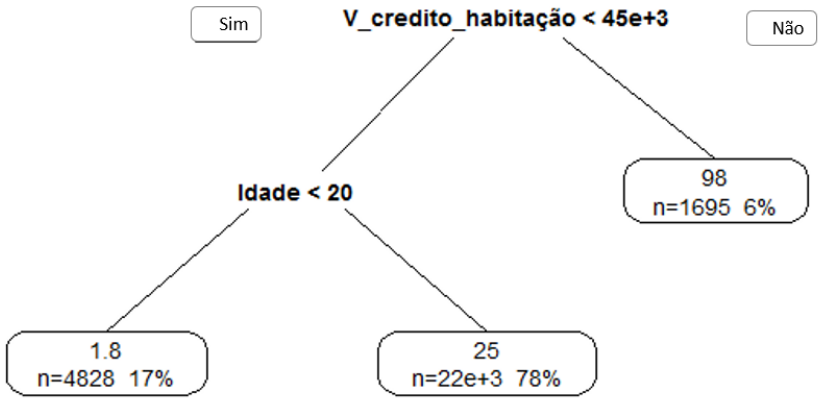


Figura 5.52: Modelo CART 6 (n - Número de clientes alocados à folha).

5.2.3.2 Construção da matriz de transição

Para o cálculo da matriz de transição, utilizou-se o mesmo mecanismo utilizado por Haenlein (2007) [22]. Os estados que foram considerados no cálculo da matriz de transição correspondem às folhas dos modelos CARTs implementadas anteriormente. Considere-se a título ilustrativo a posição p_{nn} da matriz de transição, esta posição seria calculada através da divisão entre número de clientes que num momento inicial, t_1 se encontravam no

estado n, e num momento posterior, t_2 continuam no estado n e o número total de clientes analisados. Repetindo esse procedimento para todas as possibilidades de transição entre estados, obteve-se a matriz de transição 5.1, representada por A, que será usada nesta dissertação. Salienta-se que corresponde a uma matriz $9x9$, uma vez que, as soluções ótimas obtidas através dos modelos CART para os 3 intervalos de rentabilidade do ano anterior considerados, são constituídas por 3 folhas.

$$A = \begin{bmatrix} 0.788 & 0.006 & 1.132\times10^{-4} & 0.197 & 0.003 & 0.000 & 3.772\times10^{-5} & 0.006 & 0.000 \\ 0.390 & 0.599 & 0.000 & 0.002 & 0.006 & 0.000 & 0.000 & 0.003 & 0.000 \\ 0.077 & 0.001 & 0.903 & 0.002 & 0.000 & 0.018 & 0.000 & 0.001 & 0.005 \\ 0.043 & 0.004 & 9.391\times10^{-5} & 0.851 & 0.003 & 6.260\times10^{-5} & 0.003 & 0.097 & 0.000 \\ 0.278 & 0.146 & 0.002 & 0.266 & 0.291 & 0.000 & 0.000 & 0.018 & 0.000 \\ 0.019 & 0.000 & 0.632 & 0.036 & 0.00 & 00.302 & 0.000 & 0.006 & 0.005 \\ 0.001 & 0.002 & 0.000 & 0.027 & 0.002 & 0.000 & 0.871 & 0.096 & 0.000 \\ 0.008 & 0.006 & 0.002 & 0.166 & 0.007 & 0.010 & 0.000 & 0.802 & 0.000 \\ 0.006 & 0.000 & 0.197 & 0.000 & 0.000 & 0.457 & 0.000 & 0.034 & 0.306 \end{bmatrix}. \tag{5.1}$$

Segue-se uma descrição mais pormenorizada das regras utilizadas para definir cada estado, salienta-se que as regras foram geradas pelas árvores de decisão anteriormente apresentadas:

E1: Valor do crédito à habitação $< 12\times10^3$ & Valor do crédito pessoal ≥ 2794;

E2: Valor do crédito à habitação $< 12\times10^3$ & Valor do crédito pessoal ≤ 2794;

E3: Valor do crédito à habitação $\geq 12\times10^3$;

E4: Valor do crédito à habitação $< 22\times10^3$ & Valor do crédito pessoal < 853;

E5 Valor do crédito à habitação $< 22\times10^3$ & Valor do crédito pessoal ≥ 853;

E6: Valor do crédito à habitação $\geq 22\times10^3$;

E7: Valor do crédito à habitação $< 45\times10^3$ & Idade < 20;

E8: Valor do crédito à habitação $< 45\times10^3$ & Idade ≥ 20;

E9: Valor do crédito à habitação $\geq 45\times10^3$

Os estados E1, E2 e E3 resultam da CART cuja a rentabilidade dos últimos 12 meses é superior ao terceiro quartil, os estados E4, E5 e E6 foram obtidos pela regras geradas pela CART cuja a rentabilidade dos últimos 12 meses se encontra entre o primeiro e o terceiro quartil. Por último, as regras que definem os estados E7, E8 e E9 foram geradas pela

CART cuja rentabilidade dos últimos 12 meses é inferior ao primeiro quartil. Na figura 5.53 encontra-se a representação gráfica da matriz de transição, onde estão representadas as probabilidades de transição entre estados, bem com as probabilidade de permanência no mesmo estado. As setas "circulares" indicam a probabilidade de permanência no mesmo estado. As setas "lineares" indicam a probabilidade de transição entre estados diferentes.

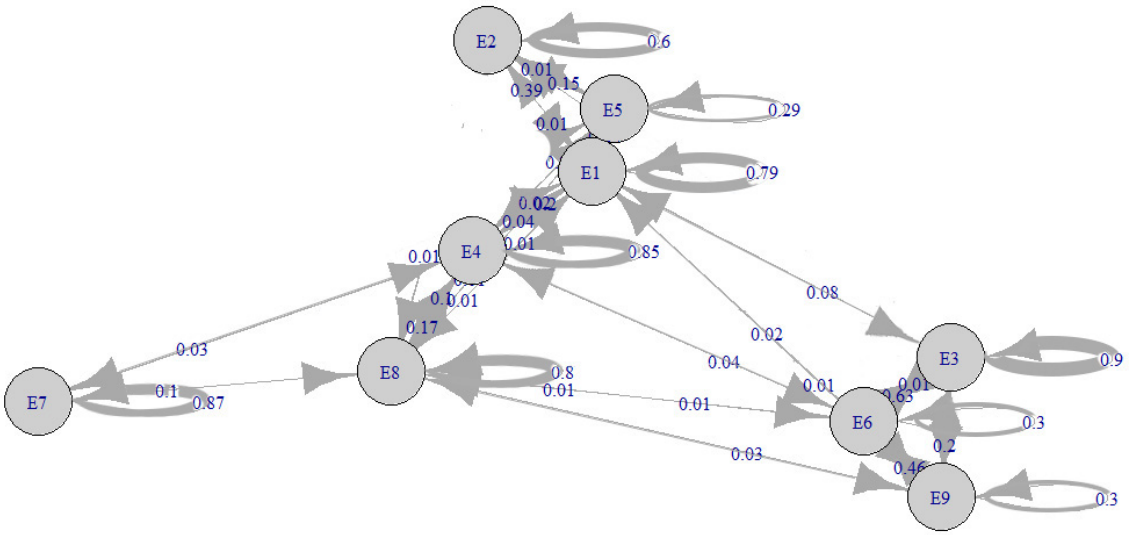


Figura 5.53: Representação gráfica da matriz de transição

5.2.3.3 Resultados e interpretação

A transição entre o estado E5 e E1, é a transição "inter" estados que tem maior probabilidade de ocorrer (27.8%). O facto de um cliente se encontrar no estado E1 significa que, apresenta valores de rentabilidade dos últimos 12 meses superiores ao terceiro quartil. Se um cliente se encontra no estado E5 significa que, o seu valor de rentabilidade dos últimos 12 meses se situa entre o primeiro e o terceiro quartil. Os critérios utilizados quer para definir o estado E1, quer o estado E5 são a posse dos produtos crédito à habitação e crédito pessoal. Assim, a passagem do estado E5 para E1 pode dever-se a continuidade da posse desses produtos, que em geral são produtos lucrativos para a instituição financeira, ou pagamento das prestações nos limites definidos.
A proximidade existente entre os estados E5 e E2 bem como, entre os estados E4 e E1 pode ter a mesma explicação da transição entre E5 e E1, já que os produtos considerados para pertencer a esses estados continuam a ser o crédito pessoal e o crédito à habitação. Verifica-se também uma elevada probabilidade de transição entre E8 e E4, esta transição pode derivar do aumento da idade do cliente em conjunto com o aumento de rentabilidade do cliente que poderá estar associada, por exemplo, à aquisição do produto crédito pessoal.
As transições de E6 para E3 e de E9 para E6, estão relacionadas com a posse do produto

crédito à habitação. A transição de E9 para E3 pode estar relacionada com amortizações do crédito. Outra justificação para esta transição é o cliente passar a utilizar o banco onde têm crédito como seu primeiro banco.
O estado E7 é o que tem maior probabilidade de permanência, o que pode estar relacionado com o facto de os clientes que se encontram nesse estado não terem atingido uma idade superior a 20 anos durante o intervalo de tempo considerado nesta análise. O estado E3 também apresenta elevada probabilidade de permanência, que pode estar relacionada com o facto de o crédito à habitação ser um produto que demora alguns anos a ser pago e enquanto o cliente o possui é considerado um cliente bastante rentável.
Salienta-se que, Haenlein (2007) optou por medir a performance das CART no "agrupamento" dos cliente e de seguida calcular a matriz de transição e os estados com base nesse pressuposto não fazendo qualquer outra avaliação à performance do modelo. Nesta dissertação apesar de também ter sido avaliada de forma rigorosa a performance das CARTs, de modo a evitar erros, também se optou por mesurar a performance após os clientes terem sido associados a determinado estado (tabela 5.24). Detalhadamente, na tabela 5.24 encontram-se as percentagem prevista (predição) e a percentagem real (validação) de clientes que se encontram em cada estado em dois momentos distintos: t_1, que corresponde à predição do *target* para um ano depois, à semelhança daquilo que foi feito nos restantes modelos e t_2, que corresponde à previsão do *target* para dois anos depois. Os resultados que se encontram na coluna predição correspondem à percentagem de clientes que o modelo previu que se se encontrassem em cada estado. Os resultados que se encontram na coluna validação correspondem à percentagem real de clientes que se encontra em cada estado.

Tabela 5.24: Resultados obtidos pelas Cadeias de Markov.

Estado	**Predição para o momento t_1(%)**	**Validação para o momento t_1(%)**	**Predição para o momento t_2(%)**	**Validação para o momento t_2(%)**
E1	15.8	19.2	18.9	18.6
E2	1.3	1.1	1.2	1.1
E3	3.5	3.1	3.2	2.8
E4	51.3	49.8	50.3	50.5
E5	0.5	0.5	0.5	0.5
E6	0.4	1.2	5.9	1.1
E7	1.0	3.7	3.4	3.1
E8	26.2	21.0	22.9	22.0
E9	0.3	0.4	0.9	0.4

5.2.4 *Multilayer Perceptron*

5.2.4.1 Parametrização e resultados

Após a "construção" da MLP procedeu-se à configuração dos parâmetros definidos no capítulo 4, secção 4.5. Para decidir quais os parâmetros que permitiam uma melhor aprendizagem por parte do modelo, de cada vez que se experimentava um novo parâmetro efectuava-se a medição do erro de treino e de teste. O *Mean Absolute Error*(MAE) foi a métrica escolhida para medir o erro enquanto se parametrizava o modelo.
O erro percentual obtido por este modelo no conjunto de teste e treino após terem sido feitas as parametrizações foi 12.88% e 11.43% respectivamente. Pelas razões descritas no capítulo 4, secção 4.5 optou-se por fazer a inicialização dos pesos e bias com a função *Randomize Weights*.
Os valores do erro de treino e teste encontram-se na tabela 5.25.

Tabela 5.25: Erro de treino e teste obtidos com a função de inicialização dos pesos e bias

Função de inicialização dos pesos e bias	**Erro de teste**	**Erro de treino**
Randomize Weights	46.0	42.2

Na tabela 5.26 encontram-se os erros de teste e treino para as funções de aprendizagem (*learning fuctions*) apresentadas no capítulo 4, secção 4.5. No modelo construído decidiu-se usar a função *Std Backpropagation* como *learning function* por ser aquela que apresentava menor valor de erro de teste e treino.

Tabela 5.26: Variação do erro de treino e teste em função da *Learning fuction*

Learning Function	**Erro de teste**	**Erro de treino**
Std Backpropagation	45.5	42.3
BackpropBatch	130.2	129.3
BackpropMomentum	46.3	42.3
BackpropChunk	45.8	42.6
BackpropWeightDecay	46.1	42.4
Rprop	47.1	44.5
Quickprop	72.8	73.9
SCG	48.9	46.3

Na tabela 5.27 encontram-se os valores do erro de treino e erro de teste para diferentes valores de *learning rate* testados. Apesar de se ter optado por usar um valor de *learning rate* baixo, 1×10^{-6}, o processo de treino foi repetido inúmeras vezes e não se verificou lentidão no mesmo.
Na tabela 5.28 encontra-se o erro de treino e de teste em função do número de épocas consideradas. No modelo construído optou-se por considerar $maxit = 150$, por esse valor ser aquele com se optem menores valores de erro de treino e erro de teste.

Tabela 5.27: Variação do erro de treino e teste em função do *learning rate*

Learning Rate	**Erro de teste**	**Erro de treino**
1×10^{-6}	46.0	42.0
1×10^{-5}	47.4	43.4
1×10^{-4}	53.2	51.7

Tabela 5.28: Variação do erro de treino e teste em função do número de épocas

Maxit	**Erro de teste**	**Erro de treino**
50	46.3	42.8
100	46.9	42.3
150	46.0	42.0
300	46.4	42.4
500	46.4	42.1
700	46.7	41.6

Uma rede pode ser constituída por várias *hidden layers*, nesta caso foram testadas redes com 1, 2, e 3 *hidden layers*. Na tabela 5.29 entram-se as várias combinações testadas para este parâmetro. Por exemplo, se o *size* definido fosse *c(5,5,5)* significava que a rede seria constituída por 3 *hidden layers* cada uma com 5 *neurons*. Os melhores resultados foram obtidos com uma rede constituída por 2 *hidden layers*, em que a primeira era formada por 15 *neurons* e a segunda por 19 *neurons*.

Tabela 5.29: Variação do erro de treino e teste em função do *size*

Size	**Erro de teste**	**Erro de treino**
c(5,9)	50.4	48.3
c(10,14)	46.3	43.8
c(15,19)	46.0	42.2
c(20,25)	46.3	42.2
c(26,30)	46.2	41.4
c(20)	49.1	46.7
c(5,5,5)	51.0	50.0
c(50)	48.3	45.7
c(10,10)	46.4	43.7
c(15,15)	46.5	42.3

Neste estudo optou-se por utilizar a opção *outputActFunc = "Act Identity"*, por apresentar um menor valor de erro de teste, como se pode verificar na tabela 5.30.

Tabela 5.30: Variação do erro de treino e teste de acordo com a função de *output* utilizada

Output	**Erro de teste**	**Erro de treino**
outputActFunc = "Act Identity"	46.0	42.0
linOut = TRUE	46.8	41.7

5.2.4.2 Modelo

Na tabela 5.31 encontra-se uma síntese dos parâmetros e respectivos valores utilizados na construção do modelo.
Depois de efectuadas todas as parametrizações o valor do erro de treino e de teste obtidos, MAE, foram 42.0 e 46.0 respectivamente. No caso do erro percentual, o valor obtido foi 12.88%.
Salienta-se que o modelo MLP era apontado na literatura como um dos modelos que apresentava melhor performance o que também se veio a verificar nesta dissertação. Dada a eficiência provada deste modelo, houve outros autores a utiliza-lo na previsão do CLV: Mauricio (2016) desenvolveu uma MLP com 1 *hidden layer*. No final, obteve um erro percentual de 1.8%. Como *input* foram utilizadas variáveis transacionais e demográficas, à semelhança daquilo que foi feito nesta dissertação. Adicionalmente, o autor considerou na sua abordagem, variáveis que estão relacionadas com os gastos e frequência associados a cada visita. Convém referir que este estudo tinha como objetivo a previsão do CLV para uma empresa de *direct selling* [36]; Sabbeh (2018) e Rathi (2011) para além dos modelos CART também construíram modelos MLP para prever o valor do CLV. O modelo de Sabbeh (2018) era constituído por uma *hidden layer* com 5 *neuron* e utilizou a "Std Backpropagation" como *learning fuction*. Obteve um erro percentual de 6.6%, o que acaba por reforçar a tese de que teria sido útil incluir no modelo variáveis relativas à propensão ao *churn*. No modelo desenvolvido por Rathi obteve-se um MAE de 2107.10.

Tabela 5.31: Parametrização da MLP

Parâmetro	**Valor utilizado**
Função de inicialiação dos pesos	*Randomize Weights*
Learning function	*Std Backpropagation*
Learning rate	1×10^{-6}
Maxit	150
Size	c(15,19)
Função de ativação do output	*outputActFunc = "Act Identity"*

5.2.5 *K-means*

O agrupamento dos clientes em *clusters*, de modo a haver um conhecimento mais aprofundado dos mesmos e assim proporcionar-lhe uma oferta mais personalizada tem sido um tema de estudo em vários projetos. Hasheminejad (2018), Hajipour (2019), Khajvand (2011) e Rabiei (2015) foram alguns autores que usaram o *k-means* nos seus projetos de modo a fazer a segmentação dos clientes à semelhança daquilo que foi feito nesta dissertação [23, 24, 29, 43]. Salienta-se que a principal diferença existente entre os modelos de *k-means* desenvolvidos por estes autores e o modelo de *k-means* desta dissertação são as variáveis utilizadas como *input*. Enquanto que na literatura se optou por utilizar as variáveis recência, frequência e valor monetário, nesta dissertação optou-se por usar variáveis transaccionais, demográficas, de fidelização e de posse.

5.2.5.1 Número de *clusters*

O método de agrupamento que foi utilizado nesta análise foi o *K-Means*, tal como descrito no capítulo 4, secção 4.7. A figura 5.54 mostra a relação entre a variabilidade intragrupo (soma dos quadrados intra-*clusters*) e o número de *clusters*. Analisando a figura concluise que, segundo o método do cotovelo, o número ideal de *clusters* é 12. Salienta-se que do ponto de vista do negócio não faria sentido desenvolver esta análise em 12 *clusters*, uma vez que, os custos associados à produção de 12 campanhas de *marketing* distintas e personalizadas seriam muito elevados. Como os decréscimos, para os casos em que se consideram mais de 5 *clusters*, tem pouca amplitude, ou seja, apresentam pouco ganho em termos de homogeneidade intra-*clusters*, decidiu-se realizar a análise de *clusters* considerando 5 *clusters*.

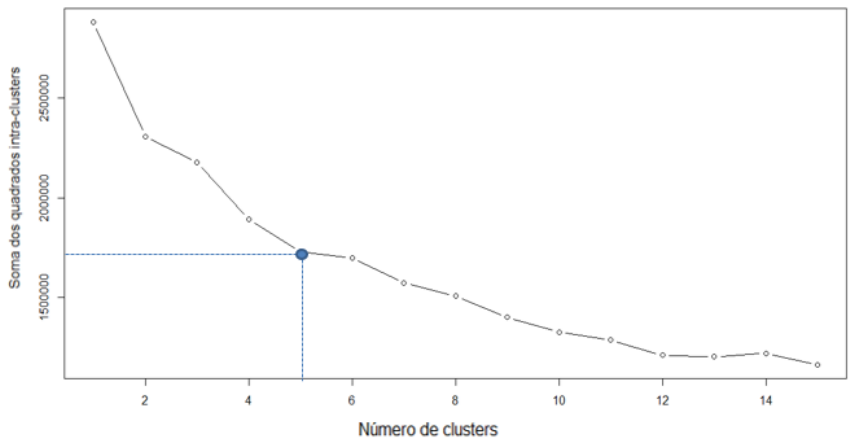


Figura 5.54: Relação entre o número de *clusters* e a variabilidade intra-*clusters* (a linha azul assinala a solução para $k = 5$)

Para agrupar os objetos do *dataset* em estudo foi feita a seleção das 20 variáveis mais importantes: a idade, o valor do seguro de vida, o montante em produtos, o valor do crédito pessoal, o valor do crédito habitação, a antiguidade do cliente no banco, o valor dos recursos a título, o valor das simulações, o valor do saldo na conta corrente, o valor do seguro multirriscos, o valor do seguro de risco não vida, a rentabilidade líquida de há 2 anos atrás, a taxa de variação da rentabilidade líquida, a taxa de variação do montante de produtos, o valor do património financeiro, o número de cartões de crédito *Visa* e *Mastercard*, a rentabilidade líquida dos últimos 12 meses, o valor de crédito vivo, o número de simulações e o montante de produtos do ano anterior. Esta selecção baseou-se nos resultados apresentados na figura 5.47.
Foram excluídas as variáveis valor do património financeiro, montante de produtos do ano anterior, valor de crédito vivo e a rentabilidade líquida dos últimos 12 meses por apresentarem correlações superiores a 70% com outras variáveis presentes no *dataset*. Na análise de *clusters*, o CLV também foi incluído como variável de *input*.
Os parâmetros adoptados na função *k-means* do R foram os seguintes: o número de centróides (*centers*), neste caso 5; a opção *nstart*, que específica o número de configurações iniciais a experimentar também igual a 5 e reporta a melhor; a opção *iter.max*, que específica o número máximo de iterações permitidas, neste caso adoptou-se o valor de *default*, 10.

5.2.5.2 Caracterização dos *clusters*

Na tabela 5.32 encontra-se a caracterização dos *clusters* atendendo às variáveis usadas na sua definição, tal como descrito no capítulo 4, secção 4.7.
Nesta secção é apresentada a caracterização detalhada de cada um dos *clusters*.

***Cluster* 1**: Neste *cluster* incluem-se aproximadamente 5% dos clientes da amostra. De acordo com os valores da tabela temos que: Este grupo é o que em média apresenta valores de seguro de vida e multirriscos mais elevados, respetivamente iguais a 319€/mês e 112€/mês. Salienta-se que, este também é um dos *cluster* que apresenta maiores valores médios para os seguros de risco não vida, cujo valor mensal é 309€, bem como no crédito à habitação e no crédito pessoal, que apresentam valores iguais a 64 474€/mês e 1 018€/mês, respetivamente.
Assim, nas campanhas de *marketing* direcionadas a este *cluster* devem ser relacionadas com seguros, uma vez que este parece ser um produto que lhes desperta bastante interesse.
Estas características sugerem tratar-se de um grupo susceptível/preocupado com a cobertura de riscos.

***Cluster* 2**: Neste *cluster* incluem-se aproximadamente 1% dos clientes da amostra. Este grupo apresenta valores médios superiores aos restantes grupos em muitas das variáveis consideradas nesta análise, nomeadamente: montante em produtos, valor

dos recursos a título, CLV, valor do saldo da conta corrente, valor dos seguros de risco não vida e rentabilidade líquida 2 anos antes, que apresentam valores médios respetivamente iguais a 105 000€/mês, 33 092€/mês, 651€/mês, 6 002€/mês, 392€/mês e 663€/mês. Também é neste *cluster* que os clientes demonstram um maior interesse pelos cartões de crédito, pois, em média, os clientes deste *cluster* possuem 0.85 cartões de crédito. No entanto, o valor médio de crédito pessoal (295€/mês) é inferior ao dos restantes *clusters*. Os clientes tem idade média 65 anos, o que significa que, este *cluster* apresenta uma idade média superior comparativamente com os outros *clusters*.

Como os clientes deste *cluster* apresentam algumas posses financeiras, não necessitando, por isso de recorrer a créditos, os produtos de investimento poderiam ser uma boa aposta para as campanhas de *marketing* direcionadas a estes clientes.

Estas características sugerem tratar-se de um grupo com capacidade financeira e interesse em rentabilizar o seu dinheiro.

***Cluster* 3**: Neste *cluster* incluem-se aproximadamente 0.1% dos clientes da amostra. Este grupo apresenta valores médios de crédito pessoal, crédito habitação e de simulações iguais a 1 1130€/mês, 76 433€/mês e 715 347€/mês respectivamente e, portanto, superiores aos restantes grupos. Neste grupo os clientes apresentam uma idade média de 39 anos, o que significa que, este *cluster* apresenta uma idade média inferior comparativamente com os outros *clusters*.

Como este grupo de clientes tem propensão para a aquisição de créditos, as campanhas de *marketing* que lhes forem encaminhadas poderiam estar direccionadas para esses produtos.

Estas características sugerem tratar-se de um grupo jovem e, talvez por isso, com pouca capacidade financeira, tendo por isso necessidade de recorrer a créditos.

***Cluster* 4**: Neste *cluster* incluem-se aproximadamente 16% dos clientes da amostra. Este grupo é um dos que apresenta valores medianos nas variáveis consideradas, e, por isso, as campanhas que lhes forem encaminhadas poderiam ter como intuito obter uma maior fidelização desses clientes.

Estas características sugerem tratar-se de um grupo com pouca vinculação ao banco.

***Cluster* 5**: Neste *cluster* incluem-se aproximadamente 78% dos clientes da amostra. Este grupo é o que apresenta valores médios inferiores em todas as variáveis com exceção do crédito pessoal e número de simulações. Os valores destas variáveis são respectivamente 609€/mês e 0.03 simulações/mês.

Para este grupo de clientes poderiam-se fazer campanhas de incentivo à aquisição de crédito pessoal, uma vez que este é o único produto que lhes desperta algum interesse.

Estas características sugerem que, este grupo apesar de apresentar pouca capacidade financeira, demonstram interesse pelo produto crédito pessoal.

Tabela 5.32: Caracterização dos *clusters* (médias mensais, por variável) obtidos através do método *K-means*

Cluster	Idade	Seguro de vida (€)	Montante em produtos (€)	Crédito pessoal (€)	Crédito habitação (€)	Número de cartões de crédito	Antiguidade
Cluster 1	54	319	41 092	1 018	64 474	0.73	21
Cluster 2	65	93	105 000	295	16 984	0.85	25
Cluster 3	39	236	26 384	1 130	76 433	0.79	12
Cluster 4	56	169	17 424	721	23 881	0.53	21
Cluster 5	51	25	1 849	609	516	0.35	17

Cluster	Recursos a título (€)	Valor das simulações (€)	Número de simulações (€)	CLV (€)	Saldo na conta corrente (€)	Seguro multirriscos (€)	Seguros de risco não vida (€)	Rentabilidade líquida 2 anos antes (€)
Cluster 1	6 104	3 586	0.03	524	1 224	112	309	320
Cluster 2	33 092	1 079	0.01	651	6 002	72	392	663
Cluster 3	788	207 080	0.16	570	832	87	257	171
Cluster 4	2 458	1 139	0.02	296	721	68	191	210
Cluster 5	211	468	0.03	107	156	14	80	79

Cluster	Taxa de variação do montante em produtos (%)	Taxa de variação da rentabilidade líquida (%)
Cluster 1	329	-2.1
Cluster 2	858	-1.1
Cluster 3	108 555	-10.4
Cluster 4	131	-0.6
Cluster 5	50	-0.2

Valor médio mais baixo registado na variável

Valor mais alto registado na variável

Capítulo 6

Conclusão

Neste capítulo serão sintetizados os principais resultados obtidos no âmbito desta dissertação.

A previsão do CLV é um tema de estudo antigo que, no entanto, continua a despertar interesse atualmente, uma vez que vivemos num mundo competitivo em que as empresas tem cada vez mais interesse em reter os clientes mais lucrativos. Assim, propôs-se como objetivo deste estudo a implementação de modelos de *Machine Learning* capazes de prever com precisão o valor do CLV dos clientes.

Após ter sido feita a revisão da literatura decidiu-se implementar os seguintes modelos: o modelo CART, por ser facilmente interpretável; o modelo *Random Forest* por apresentar uma complexidade superior ao modelo CART e por isso, apresentar teoricamente, melhores resultados; o modelo MLP que pela sua complexidade apresenta, em teoria bons resultados, dada a sua facilidade em descobrir padrões ocultos e as Cadeias de Markov por permitirem determinar a probabilidade de transição entre estados (conjunto de regras geradas com recurso a árvores de decisão, que definem a "situação" de determinado cliente com base nos seus valores de crédito à habitação e crédito pessoal em determinado instante). Dos modelos implementados aquele que teve melhor performance foi o *Random Forest*, com um erro percentual de 5.98%, seguindo-se os modelos MLP e CART, com erros percentuais respetivamente iguais a 12.88% e 17.25%.

Assumindo o modelo CART as variáveis mais importantes para a previsão foram: a rentabilidade líquida dos últimos 12 meses, o valor de crédito pessoal, valor reclamado e o valor de poupança à habitação, uma vez que estas variáveis apresentaram ganhos de informação superiores a 10%. No caso do modelo *Random Forest*, as variáveis que mais se destacaram na previsão do CLV foram: a rentabilidade líquida dos últimos 12 meses, o saldo da conta corrente e a taxa de variação do montante em produtos *cross-selling*,

apresentando percentagens de IncMSE superiores ou iguais a 20%.
Da análise das Cadeias de Markov concluiu-se que os fatores que mais influenciam a previsão do CLV são os valores do crédito pessoal e do crédito à habitação, influenciando a definição dos estados e a probabilidade de transição. Além disso, verificou-se também que os clientes que apresentam uma idade inferior a 20 anos, são aqueles que apresentam uma maior probabilidade de permanência no mesmo estado.
Na literatura foi referida, por diversas vezes, a importância das empresas terem uma estratégia de retenção de clientes "eficaz", uma vez que é mais barato reter os clientes existentes do que adquirir novos clientes. Com o intuito de compreender melhor os interesses dos diferentes grupos de clientes, foi também realizada uma análise de *clusters* utilizando o método *K-means*. Nesta análise, foram incluídas as variáveis que, de acordo com o modelo *Random Forest* eram as mais importantes para a previsão do CLV do cliente, bem como o seu valor de CLV. No final obtiveram-se 5 grupos distintos. Com esta informação, a empresa poderá desenhar campanhas de *marketing* adequadas aos interesses do cliente, tendo também em conta o seu valor de CLV.
Assim conclui-se que atualmente já é possível fazer a previsão do valor de CLV dos clientes com eficácia bem como desenhar estratégias de *marketing* personalizadas com o objetivo de reter clientes tendo por base o seu valor de CLV.
Salienta-se que este trabalho também contribui para a evolução do estado da arte uma vez que, não tinham sido anteriormente reportadas análises de *clusters* em que as variáveis de *input* utilizadas não fossem a recência, frequência e valor monetário.

Referências

[1] H. Aeron, A. Kumar e J. Moorthy. "Data mining framework for customer lifetime value-based segmentation". Em: *Journal of Database Marketing and Customer Strategy Management* 19.1 (2012), pp. 17–30. ISSN: 17412439. DOI: 10.1057/dbm.2012.1.

[2] J. Anton. "The past, present and future of customer access centers". Em: *International Journal of Service Industry Management* 11.2 (2000), pp. 120–130. ISSN: 09564233. DOI: 10.1108/09564230010323534.

[3] M Ayoubi. "Customer Segmentation Based on CLV Model and Neural Network". Em: *International Journal of Computer Science Issues* 13.2 (2016), pp. 31–37. ISSN: 16940814. DOI: 10.20943/01201602.3137.

[4] E. N. G. M. L. D. R. R. Berger P. e. Venkatesan R. "From Customer Lifetime Value to Shareholder Value". Em: *Journal of Service Research* (2006), pp. 156–167.

[5] P. D. Berger e N. I. Nasr. "83_Calc_CustLifetimeValue_2". Em: (1998), pp. 17–30.

[6] C. Bergmeir e J. M. Benítez. "Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS". Em: *Journal of Statistical Software* 46.7 (2012), pp. 1–26. URL: http://www.jstatsoft.org/v46/i07/.

[7] R. C. Blattberg e J. Deighton. "Manage Marketing by". Em: *Harvard Business Review* July-Augus (1996), pp. 205–213. DOI: 10.1142/9789814287067{_}0013.

[8] R. N. Bolton, K. N. Lemon e P. C. Verhoef. "The theoretical underpinnings of customer asset management: A framework and propositions for future research". Em: *Journal of the Academy of Marketing Science* 32.3 (2004), pp. 271–292. ISSN: 00920703. DOI: 10.1177/0092070304263341.

[9] L. Breiman. *Random Forest. Machine learning 45, 1, 5-32.* 2001.

[10] D. Budale e D. Mane. "Predictive Analytics in Retail Banking". Em: *International Journal of Engineering and Advanced Technology* 2.5 (2013), pp. 508–510.

[11] D. Collings e N. Baxter. "Valuing customers". Em: *BT Technology Journal* 23.3 (2005), pp. 24–29. ISSN: 13583948. DOI: 10.1007/s10550-005-0027-0.

[12] M. Crowder, D. J. Hand e W. Krzanowski. "On optimal intervention for customer lifetime value". Em: *European Journal of Operational Research* 183.3 (2007), pp. 1550–1559. ISSN: 03772217. DOI: 10.1016/j.ejor.2006.08.062.

[13] M. Dash e P. W. Koot. "Feature Selection for Clustering". Em: *Encyclopedia of Database Systems*. Vol. 1805. New York, NY: Springer New York, 2016, pp. 1–8. ISBN: 3540673822. DOI: 10.1007/978-1-4899-7993-3{_}613-2. URL: http://link.springer.com/10.1007/978-1-4899-7993-3_613-2.

[14] J. W. (ed.) *Encyclopedia of Business Analytics and Optimization*. 1ª ed. IGI Global, 2014. ISBN: 1466652020,9781466652026. URL: http://gen.lib.rus.ec/book/index.php?md5=990d8557404f9f054e407100338c37f0.

[15] Y. Ekinci, N. Uray e F. Ülengin. "A customer lifetime value model for the banking industry: a guide to marketing actions". Em: *European Journal of Marketing* 48.3/4 (2014), pp. 761–784. ISSN: 0309-0566. DOI: 10.1108/ejm-12-2011-0714.

[16] M. EsmaeiliGookeh e M. Tarokh. "Customer Lifetime Value Models: A literature survey". Em: *International Journal of Industrial Engineering and Production Management* 24.4 (2013), pp. 317–336. DOI: 2008-4889.

[17] Euribor. *Euribor*. 2020. URL: https://www.euribor-rates.eu/ (acedido em 09/02/2020).

[18] P. S. Fader, B. G. Hardie e K. Jerath. "Estimating CLV using aggregated data: The Tuscan Lifestyles case revisited". Em: *Journal of Interactive Marketing* 21.3 (2007), pp. 55–71. ISSN: 15206653. DOI: 10.1002/dir.20085.

[19] P. S. Fader, B. G. Hardie e J. Shang. "Customer-base analysis in a discrete-time noncontractual setting". Em: *Marketing Science* 29.6 (2010), pp. 1086–1108. ISSN: 07322399. DOI: 10.1287/mksc.1100.0580.

[20] M. Galal, G. Hassan e M. Aref. "Developing a personalized multi-dimensional framework using business intelligence techniques in banking". Em: *ACM International Conference Proceeding Series* 09-11-May- (2016), pp. 21–27. DOI: 10.1145/2908446.2908488.

[21] S. Gupta, D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, N. Ravishanker e S. Sriram. "Modeling customer lifetime value". Em: *Journal of Service Research* 9.2 (2006), pp. 139–155. ISSN: 10946705. DOI: 10.1177/1094670506293810.

[22] M. Haenlein, A. M. Kaplan e A. J. Beeser. "A Model to Determine Customer Lifetime Value in a Retail Banking Context". Em: *European Management Journal* 25.3 (2007), pp. 221–234. ISSN: 02632373. DOI: 10.1016/j.emj.2007.01.004.

[23] B. Hajipour e M. Esfahani. "Delta model application for developing customer lifetime value". Em: *Marketing Intelligence and Planning* 37.3 (2019), pp. 298–309. ISSN: 02634503. DOI: 10.1108/MIP-06-2018-0190.

[24] S. M. H. Hasheminejad e M. Khorrami. "Data mining techniques for analyzing bank customers: A survey". Em: *Intelligent Decision Technologies* 12.3 (2018), pp. 303–321. ISSN: 18758843. DOI: 10.3233/IDT-180335.

[25] M. B. Hosseni e M. J. Tarokh. "Customer Segmentation Using CLV Elements". Em: *Journal of Service Science and Management* 04.03 (2011), pp. 284–290. ISSN: 1940-9893. DOI: 10.4236/jssm.2011.43034.

[26] H. Hwang, T. Jung e E. Suh. "An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry". Em: *Expert Systems with Applications* 26.2 (2004), pp. 181–188. ISSN: 09574174. DOI: 10.1016/S0957-4174(03)00133-7.

[27] D. Jain e S. S. Singh. "Customer lifetime value research in marketing: A review and future directions". Em: *Journal of Interactive Marketing* 16.2 (2002), pp. 34–46. ISSN: 10949968. DOI: 10.1002/dir.10032.

[28] M. S. Kahreh, M. Tive, A. Babania e M. Hesan. "Analyzing the Applications of Customer Lifetime Value (CLV) based on Benefit Segmentation for the Banking Sector". Em: *Procedia - Social and Behavioral Sciences* 109.October 2015 (2014), pp. 590–594. ISSN: 18770428. DOI: 10.1016/j.sbspro.2013.12.511.

[29] M. Khajvand e M. J. Tarokh. "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context". Em: *Procedia Computer Science* 3 (2011), pp. 1327–1332. ISSN: 18770509. DOI: 10.1016/j.procs.2011.01.011. URL: http://dx.doi.org/10.1016/j.procs.2011.01.011.

[30] Kumar. *vkclv*. 2020. URL: http://www.vkclv.com/ (acedido em 09/02/2020).

[31] V. Kumar. *Customer lifetime value - The path to profitability*. Vol. 2. 1. 2007, pp. 1–96. ISBN: 1700000004. DOI: 10.1561/1700000004.

[32] V. Kumar e M. George. "Measuring and maximizing customer equity: a critical analysis". Em: *Journal of the Academy of Marketing Science* 35.2 (2007), pp. 157–171. ISSN: 00920703. DOI: 10.1007/s11747-007-0028-2.

[33] V. Kumar, I. D. Pozza e J. Ganesh. "Revisiting the satisfaction-loyalty relationship: Empirical generalizations and directions for future research". Em: *Journal of Retailing* 89.3 (2013), pp. 246–262. ISSN: 00224359. DOI: 10.1016/j.jretai.2013.02.001. URL: http://dx.doi.org/10.1016/j.jretai.2013.02.001.

[34] A. Liaw e M. Wiener. "Classification and Regression by randomForest". Em: *R News* 2.3 (2002), pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

[35] M. Lycett e A. Marshan. "Modeling connected customer lifetime value (CCLV) in the banking domain". Em: *AMCIS 2017 - America's Conference on Information Systems: A Tradition of Innovation* 2017-Augus.August (2017).

[36] A. P. Mauricio, J. M. M. Payawal, M. A. D. Cueva e V. C. Quevedo. "Mining Technique in a Direct Selling Company". Em: (2016).

[37] S. Munk. "Gadgets". Em: *Engineering and Technology* 7.12 (2012), pp. 88–89. ISSN: 17509637. DOI: 10.1049/et.2012.1230.

[38] M. B. Mzoughia, S. Borle e M. Limam. "A MCMC approach for modeling customer lifetime behavior using the COM-Poisson distribution". Em: *Applied Stochastic Models in Business and Industry* 34.2 (2018), pp. 113–127. ISSN: 15264025. DOI: 10.1002/asmb.2276.

[39] B. Noori. "An Analysis of Mobile Banking User Behavior Using Customer Segmentation". Em: *International Journal of Global Business* 8.December (2015), pp. 55–64.

[40] r. *information$_g$ain*. 2020. URL: https://www.rdocumentation.org/packages/FSelectorRcpp/versions/0.3.3/topics/information_gain (acedido em 09/02/2020).

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: https://www.R-project.org/.

[42] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2018. URL: https://www.R-project.org/.

[43] A. Rabiei. "Integrating RFM and Classification for Response Modeling Based on Customer Lifetime Value". Em: *Cumhuriyet Science Journal* 36.4 (2015), pp. 246–253. ISSN: 1300-1949. DOI: 10.17776/csj.05548.

[44] Rathi. "Customer Lifetime Value Measurement using Machine Learning Techniques". Em: (2011).

[45] W. Reinartz, J. S. Thomas e V. Kumar. "Balancing acquisition and retention resources to maximize customer profitability". Em: *Journal of Marketing* 69.1 (2005), pp. 63–79. ISSN: 00222429. DOI: 10.1509/jmkg.69.1.63.55511.

[46] W. J. Reinartz e V. Kumar. "On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing". Em: *Journal of Marketing* 64.4 (2000), pp. 17–35. ISSN: 00222429. DOI: 10.1509/jmkg.64.4.17.18077.

[47] P. Romanski e L. Kotthoff. *FSelector: Selecting Attributes*. R package version 0.21. 2016. URL: https://CRAN.R-project.org/package=FSelector.

[48] R. T. Rust, C. Moorman e G. Bhalla. "Rethinking marketing". Em: *Harvard Business Review* 88.1-2 (2010). ISSN: 00178012. DOI: 10.4324/9781351283083-3.

[49] S. F. Sabbeh. "Machine-learning techniques for customer retention: A comparative study". Em: *International Journal of Advanced Computer Science and Applications* 9.2 (2018), pp. 273–281. ISSN: 21565570. DOI: 10.14569/IJACSA.2018.090238.

[50] *SAS 9.2 (SAS Institute, Cary NC)*. 2008.

[51] P. Schmitt, B. Skiera e C. Van Den Bulte. "Referral programs and customer value". Em: *Journal of Marketing* 75.1 (2011), pp. 46–59. ISSN: 00222429. DOI: 10.1509/jmkg.75.1.46.

[52] J. Shao, X. Li e W. Liu. "The application of AdaBoost in customer churn prediction". Em: *Proceedings - ICSSSM'07: 2007 International Conference on Service Systems and Service Management* 00 (2007). DOI: 10.1109/ICSSSM.2007.4280172.

[53] F. Shirazi e M. Mohammadi. "A big data analytics model for customer churn prediction in the retiree segment". Em: *International Journal of Information Management* 48.October (2019), pp. 238–253. ISSN: 02684012. DOI: 10.1016/j.ijinfomgt.2018.10.005. URL: https://doi.org/10.1016/j.ijinfomgt.2018.10.005.

[54] G. A. Spedicato. "Discrete Time Markov Chains with R". Em: *The R Journal* (jul. de 2017). R package version 0.6.9.7. URL: https://journal.r-project.org/archive/2017/RJ-2017-036/index.html.

[55] H. K. Stahl, K. Matzler e H. H. Hinterhuber. "Linking customer lifetime value with shareholder value". Em: *Industrial Marketing Management* 32.4 (2003), pp. 267–279. ISSN: 00198501. DOI: 10.1016/S0019-8501(02)00188-8.

[56] N. Sun, J. G. Morris, J. Xu, X. Zhu e M. Xie. "ICARE: A framework for big data-based banking customer analytics". Em: *IBM Journal of Research and Development* 58.5-6 (2014), pp. 1–9. ISSN: 21518556. DOI: 10.1147/JRD.2014.2337118.

[57] A. Susoykina. *Customer Lifetime Value Using Statistical Modeling*. 2011.

[58] T. Therneau e B. Atkinson. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15. 2019. URL: https://CRAN.R-project.org/package=rpart.

[59] A. Vieira e A. Sehgal. "How banks can better serve their customers through artificial techniques". Em: *Digital Marketplaces Unleashed* (2017), pp. 311–326. DOI: 10.1007/978-3-662-49275-8{_}31.

[60] L. Wu, L. Liu e J. Li. "Evaluating customer lifetime value for customer recommendation". Em: *2005 International Conference on Services Systems and Services Management, Proceedings of ICSSSM'05* 1 (2005), pp. 138–143. DOI: 10.1109/ICSSSM.2005.1499450.

[61] F. Yoseph e M. Heikkila. "Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method". Em: *Proceedings - International Conference on Machine Learning and Data Engineering, iCMLDE 2018* Clv (2019), pp. 77–82. DOI: 10.1109/iCMLDE.2018.00029.

[62] Y. Zhang, Y. Ma e X. Yang. "Multi-label feature selection based on mutual information". Em: *ICNC-FSKD 2018 - 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery* 27.8 (2018), pp. 1379–1386. DOI: 10.1109/FSKD.2018.8687220.

APÊNDICE A

Cadeias de Markov: Definição dos estados utilizando a rentabilidade do ano anterior

CARTs para valores de rentabilidade do ano anterior superiores ao terceiro quartil (Grupo 1)

Neste grupo foram gerados 3 modelos CART, que irão ser designados por CART 10, CART 11 e CART 12. Na tabela A.1 apresenta-se o erro de teste (calculado utilizando o *Mean Absolute Error*), o número de folhas e o parâmetro de complexidade (cp) associado a cada uma das CARTs.

A CART 11 apresenta um erro superior ao das restantes CARTs. Os valores dos erros obtidos para as CART 10 e CART 12 são muito idênticos, por esse motivo, o critério de escolha entre estas árvores foi o número de folhas. A CART 12 (figura A.1) é a que apresenta menor número de folhas e por esse motivo foi a escolhida para definir os estados para este grupo.

Tabela A.1: Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp)

CART	**Erro de teste**	**Número de folhas**	**Parâmetro de complexidade (cp)**
CART 10	62.0	6	1.0×10^{-2}
CART 11	69.0	3	4.7×10^{-2}
CART 12	63.8	5	3.0×10^{-3}

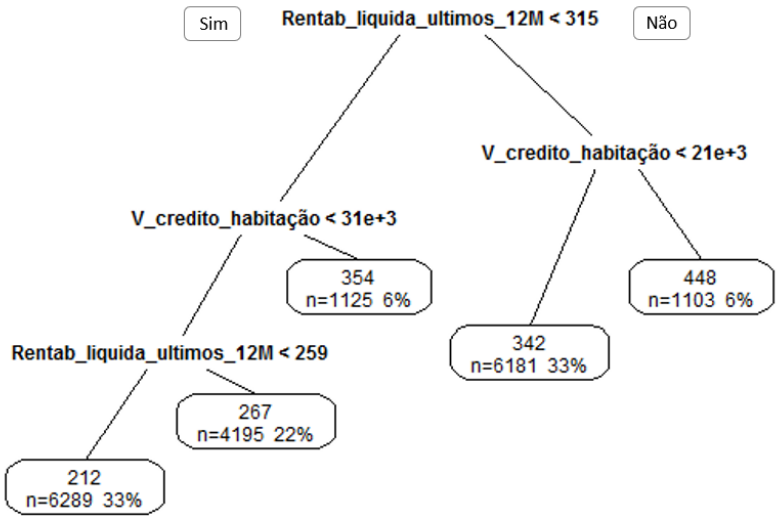


Figura A.1: Modelo CART 12 (n - Número de clientes alocados à folha).

CARTs para valores de rentabilidade do ano anterior entre o primeiro e o terceiro quartil (Grupo 2)

Neste grupo foram gerados 2 modelos CART, que irão ser designados por CART 12 e CART 13. Na tabela A.2 encontra-se a variação do erro de teste (*Mean Absolute Error*) em função do parâmetro de complexidade e número de folhas. A CART que será considerada para definir os estado para este subconjunto será a CART 13 (figura A.2) porque apresenta um erro menor.

Tabela A.2: Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp)

CART	**Erro de teste**	**Número de folhas**	**Parâmetro de complexidade (cp)**
CART 13	18.9	5	3.0×10^{-2}
CART 14	22.6	4	4.0×10^{-2}

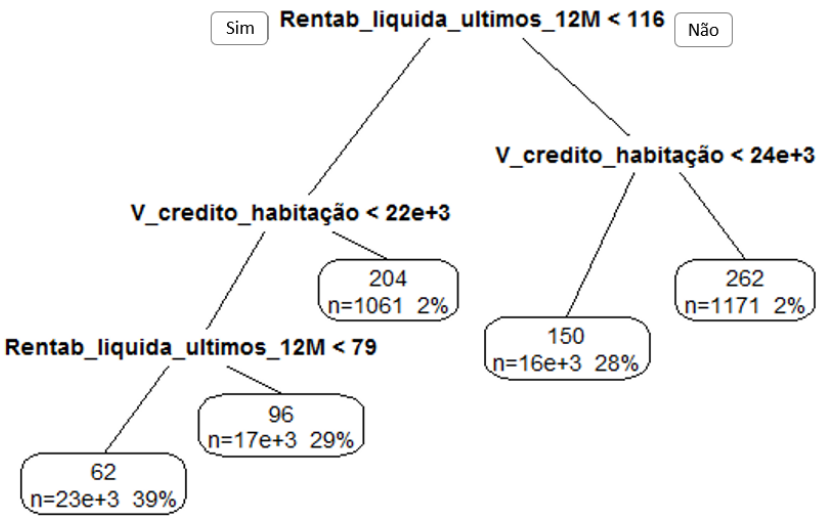


Figura A.2: Modelo CART 13 (n - Número de clientes alocados à folha).

CARTs para valores de rentabilidade do ano anterior inferiores ao primeiro quartil (Grupo 3)

Neste grupo foram gerados 2 modelos CART, que irão ser designados por CART 15 e CART 16. Na tabela A.3 encontra-se a variação do erro de teste (*Mean Absolute Error*) em função do número de folhas e parâmetro de complexidade. Neste caso, a CART 16 (figura A.3) é a melhor escolha por apresentar menores valores de erro e menos folhas.

Tabela A.3: Variação do *Mean Absolute Error* (erro de teste) em função do número de folhas e parâmetro de complexidade (cp)

CART	Erro de teste	Número de folhas	Parâmetro de complexidade (cp)
CART 15	14.8	5	1.0×10^{-2}
CART 16	14.2	3	3.0×10^{-2}

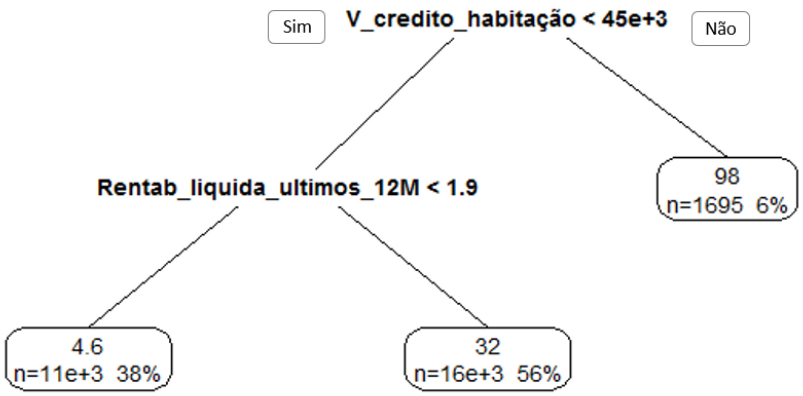


Figura A.3: Modelo CART 16 (n - Número de clientes alocados à folha).